CATTELL-HORN-CARROLL (CHC) THEORY AND MEAN DIFFERENCE IN INTELLIGENCE SCORES

By

OLIVER WAYNE EDWARDS

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA
2003

ACKNOWLEDGMENTS

First of all, I would like to express me deepest gratitude to my major professor, Dr. Thomas Oakland. His guidance and assistance were extremely instrumental in the completion of this project. I also am very grateful to my supervisory committee members, Drs. Nancy Waldron, M. David Miller, and W. Max Parker, for their insightful and incisive comments. Their knowledge and assistance were indispensable in the completion of this work. Additionally, I thank Drs. Richard Woodcock and Kevin McGrew for their permission to use the WJ-III data. Finally, I appreciate very much the innumerable others who assisted me in my educational journey.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS ii

LIST OF TABLES v

LIST OF FIGURES vii

ABSTRACT vii

CHAPTER

1 INTRODUCTION 1

Use of Intelligence Tests 1
Statement of the Problem 2
Increases in IQ Over Time 4
Historical Origins of Intelligence Testing 6
Theories of Intelligence 8
Spearman's g 8
Thurstone's Primary Mental Abilities 9
Cattell and Horn: Fluid and Crystallized Intelligence 10
Carroll's Three-Stratum Theory of Cognitive Abilities 10
Cattell-Horn-Carroll Theory of Intelligence 11
Purpose of the Study 16

2 REVIEW OF THE LITERATURE 21

The Development of Intelligence 21
Pros and Cons of Intelligence Testing 22
The Cultural Influence on IQ 24
Case Law, Cultural Bias, and Intelligence Testing 25
Special Education Eligibility and Intelligence Testing 26
Overrepresentation of Minorities in Special Education 29
Test Bias 33
Recent Concepts of Test Validity 36
Social Validity 37
Statement of Hypotheses 39

page

3 METHODS 44

Participants 44
Instrumentation 45
Test Reliability 47
Test Validity 51
Test Fairness 52
Factor Analysis 53
Procedures 54
Methodology 54

4 RESULTS 57

Principal Component Factor Analysis 57
MANOVA 57
Effect Size Test for Large Samples 58
Sigma Difference Test 58
Correlations Between General Intelligence and Achievement 60

5 DISCUSSION 73

Smaller Difference on Broad Factors than on g 74
Similar Factor Structures for Both Groups 75
Significance of g 76
Consequential Validity Perspective 78
Test Selections and Administration 79
The Importance of Intelligence Tests 82
Supplementing or Supplanting Intelligence Tests? 82
Equalizing Outcomes or Equalizing Opportunities 84

LIST OF REFERENCES 87

BIOGRAPHICAL SKETCH 96

LIST OF TABLES

Table		page
1-1	Carroll's Stratum I: Each Narrow Ability is Subsumed Under a Broad Ability	13
2-1	Percentage of student ages 6 through 21 Served by Disability and Race/ethnicity in the 1998-1999 School Year	32
3-1	Reliabilty Statistics for the WJ-III Tests of Cognitive and Achievement	48
3-2	Comparison of Fit of WJ-III CHC Broad Model Factor Structure with Alternative Models in the Age 6 to Adult Norming Sample	49
3-3	Confirmatory Factor Analysis Broad Model, g-loadings – Age 6 to Adult Norming Sample	50
4-1	WJ-III Cognitive and Achievement Batteries Codes	62
4-2	Box's Test of Equality of Covariance Matrices – Homogeneity of the Variance	63
4-3	Bartlett's Test of Sphericity	64
4-4	Multivariate Tests of Significance Effect for Group	65
4-5	Levene's Test of Equality of Error Variances	66
4-6	Univariate Tests	67
4-7	Sigma Difference - Direct Comparison of Changes in Effect Size for the GIA and Each Stratum II Subtest	68
4-8	Principal Component Matrix	69
4-9	Descriptive Statistics – Caucasian Americans and African-Americans	70

Table page

4-10 Pearson Correlations Between General Intelligence and Academic Achievement for African-Americans and Caucasian-Americans 71

4-11 Fisher Z Transformation: z-test for Independent Correlations between Caucasian-Americans and African-Americans for General Intelligence and Academic Achievement 72

LIST OF FIGURES

Figure page

1-1 Carroll's Strata II and III 12

3-1 WJ-III Tests of Cognitive Abilities as it Represents CHC Theory 46

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

CATTELL-HORN-CARROLL (CHC) THEORY
AND MEAN DIFFERENCE IN INTELLIGENCE SCORES

By

Oliver W. Edwards

May 2003

Chair: Thomas D. Oakland
Major Department: Educational Psychology

The use of intellectual and other forms of psychological and mental tests with students who differ culturally, linguistically, or racially is subject to substantial controversy. Professionals responsible for the assessments of culturally different children frequently are uncertain which test instruments provide the most valid, relevant, and equitable results. Research studies indicate mean IQs for some racial/ethnic groups are significantly lower than mean IQs for Caucasians. Some believe IQ differences among racial/ethnic groups suggest the tests unfairly favor one group over another and evidence of group differences indicate intelligence tests are biased against lower performing groups. They further contend intelligence testing influences the disproportionate representation of minority students in special education. Most intelligence test developers currently do not provide information about mean IQ differences by racial/ethnic groups. The Woodcock-Johnson III Cognitive and Achievement Batteries were used to compare the mean score differences of the distributions between African-

Americans and Caucasian-Americans. The factor structures of the two groups were also analyzed. In light of the Spearman-Jensen hypothesis and Cattell-Horn-Carroll theory, the mean IQ difference between African-Americans and Caucasian-Americans were hypothesized to be smaller on the Woodcock-Johnson III than on other frequently used measures of intelligence. The results reveal mean IQ differences between Caucasian-Americans and African-Americans are smaller on the Woodcock-Johnson III than on other measures of intelligence. African-Americans obtain lower mean IQs than Caucasian-Americans. The factor structures of the two groups do not differ. Judgments regarding test selection and administration when mean IQ differences occur between two statistically sound instruments will influence educational decision-making and disproportionate representation of minorities in special education. All else being equal, an intelligence test with a smaller disparate mean difference between subgroups is the test that possesses less consequential bias and provides the most relevant and equitable results.

CHAPTER 1
INTRODUCTION

Use of Intelligence Tests

The use of intellectual and other forms of psychological and mental tests with students who differ culturally, linguistically, or racially is subject to substantial controversy. Professionals responsible for the assessments of culturally different children frequently are uncertain which test instruments provide the most valid, relevant, and equitable results. Interest in providing fair and equitable mental test results extends back several decades, but what is considered fair and equitable changes as the values in our culture change (Oakland, 1976; Oakland & Laosa, 1976).

In previous years, intelligence test developers (cf. the early editions of Wechsler and Standford-Binet scales) often provided test users information about mean score differences for children who differed by socioeconomic status (SES), primary language, parents' educational level, gender, and race. Information about standard score differences among racial/ethnic groups helps determine the relevance and usefulness of an intelligence test with different groups. It also encourages evaluation of the test to ascertain whether it may be biased. This process changed over the past decade, and data about mean standard score differences currently are not provided.

Differences in intelligence scores for racial/ethnic groups are considered important, in part, since tests are statistically structured to distinguish between individuals, and groups, because groups are aggregates of individuals. Intelligence tests are designed carefully and deliberately to produce score variance (Wesson, 2000). The

generation of a broad range of individual scores permits psychologists to acquire knowledge and make judgments about, between, and within group differences. This knowledge allows for the interpretation of the distribution of scores that lead to various decisions (e.g., eligibility for placement in special education and gifted programs).

Statement of the Problem

Mean IQs for some minority racial/ethnic groups are significantly lower than mean IQ for Caucasians (Jensen, 1980). The hierarchical order of intelligence test scores traditionally places Asian Americans at the top followed by Caucasian-Americans, Hispanic-Americans, and African-Americans (Jensen, 1980; Onwuegbuzie & Daley, 2001; Wesson, 2000). On average, and when unadjusted for differences in SES, Asian-Americans score approximately three points higher than Caucasian-Americans, African-Americans score approximately 15 points lower than Caucasian-Americans, and Hispanic-Americans score somewhere in between the latter two groups (Herrnstein & Murray, 1994; Onwuegbuzie & Daley, 2001). The 15-point (i.e., one standard deviation) difference detected between African-Americans and Caucasian-Americans was reported in 1932 in the United States during the development of the Army Alpha and Beta tests administered to recruits during World War I (Loehlin, Lindzey, & Spuhler, 1975). A meta-analytic study of 156 independent data sets regarding racial/ethnic IQ differences revealed an overall average difference of 16.2 points (Jensen, 1998). To ease in recall, scholars have used a 15-point difference (or one standard deviation on most intelligence tests) to reference the traditional mean IQ differences between racial/ethnic groups. The fairly consistent finding of mean IQ differences between African-Americans and Caucasian-Americans has generated considerable debate, historically and currently.

Most intelligence test developers currently do not provide information about mean IQ difference by racial/ethnic groups. The withholding of this information may be to avoid controversy and to show social sensitivity. That is, test developers may be apprehensive about appearing insensitive to some minority groups when publishing data that reflect negatively on said group. Some believe IQ differences among racial/ethnic groups suggest the tests unfairly favor one group over another and evidence of group differences indicate intelligence tests are biased against lower performing groups (Gould, 1996; Kamin, 1974; Ogbu, 1994; Onwuegbuzie & Daley, 2001). Test developers may wish to appear in support of an egalitarian ideal that maintains all subgroups within a population perform somewhat equally on measures of various traits.

Problematically, however, without data on mean IQs of various racial/ethnic groups, test performance must be interpreted in light of a common norm despite possible IQ differences among racial/ethnic groups. A common norm does not provide information specific to cultural and racial/ethnic groups. Exclusive utilization of a common norm when interpreting intelligence test scores can lead to disproportionate placement of subgroups in a variety of educational programs.

It is a challenge to interpret test scores appropriately for all examinees (Scheuneman & Oakland, 1998). The capability of interpreting test results from a variety of points of reference assists scholars to better understand and apply intelligence test scores of minority subgroups. Of course, test users and consumers of test information should be informed as to which reference point (e.g., which norm) was used and why it was chosen (Sattler, 2001).

The availability of data on mean IQ differences among racial/ethnic groups makes information accessible to tests users as to which tests are socially valid (as described

below) and most fairly reflect the intellectual functioning of minority groups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Messick, 1995). Failure to provide these data limits test users' ability to make informed choices about which intelligence tests are most equitable and appropriate to use.

For test scores to be considered socially valid, they need to be interpreted in view of the test's statistical validity as well as the value implications of the meaning of the score (e.g., are intelligence tests measures of past achievement or of ability for future achievement?). In addition, tests need to be interpreted considering the resultant social and educational consequences (e.g., special education placement) of score use (DeLeon, 1990; Messick, 1995).

Increases in IQ Over Time

Discourse on IQ differences should reference substantial increases in intelligence scores during the last 60 years. Scores on measures of intellectual functioning have risen, and in some cases risen rather sharply, during this period (Flynn, 1999; Neisser, 1998). Analysis of intelligence data from several countries (e.g., Belgium, France, Norway, Denmark, Germany, Austria, Switzerland, Japan, China, Israel, Brazil, Canada, Britain, and the United States of America) found without exception large gains in IQs over time (Flynn, 1998). The pattern of gains corresponds with the worldwide move from an agriculture-based economy to industrialization (Flynn, 1987, 1994, 1999; Raven, Raven, & Court, 1993).

Average IQs have risen by about three points a decade during the last 50 years (Flynn, 1999). These IQ gains across decades, referred to as the "Flynn effect," provide evidence that gains in average IQ are part of a persistent and perhaps universal

phenomenon (Flynn, 1999; Herrnstein & Murray, 1994). Gains are most dramatic on tests that assess a general factor, g, of intelligence. One of the best examples of an intelligence test that primarily measures g is the Raven's Progressive Matrices (Jensen, 1980). On the Raven's, one identifies the missing parts of patterns that are postulated as readily perceived by people from the majority of cultures (Flynn, 1998).

Research with the Raven's Progressive Matrices is particularly relevant because of the finding that, on tests such as the Raven's, IQ differences between African-Americans and Caucasian-Americans exceed 15 points (Jensen, 1980). The Raven's Progressive Matrices is considered to be the best-known, most extensively researched, and most widely used culture-reduced test of intelligence (Jensen, 1980). Many scholars believe the test measures g and little else and may be the most reliable measure to identify intellectually able children from impoverished backgrounds (Jensen, 1980).

However, Raven's scores may be highly influenced by environmental variables. To illustrate, all 18-year-old males in the Netherlands take an adaptation of the Raven's upon entrance into the military. Data available from this population reveal the mean scores of those tested between 1952 and 1982 rose 21 IQ points. Genetic changes within populations do not occur in such a short time span (Flynn, 1999). Therefore, the increase in Raven's IQs could be a function of changes in the environment (Neisser, 1998). Current geometric rates of change in society (e.g., the acquisition of information as a result of computers and the Internet) may lead to concomitant changes in population IQs and, important to this study, changes in subgroup IQ differences. The unknown factors producing secular IQ gains over generations may also occur within generations and lead to IQ differences among subgroups (Flynn, 1987). Thus, the finding of substantial changes in population IQs over time raises the question as to whether the historically

observed pattern of mean IQ differences among racial/ethnic groups also shows substantial change.

Historical Origins of Intelligence Testing

Empirical support for the theoretical basis of intelligence tests essentially began with the development of factor analysis (Ittenbach, Esters, & Wainer, 1997). The historical antecedents for factor analysis originated with the work of Galton who developed many of the quantitative devices utilized in psychometry (e.g., the bivariate scatter diagram, regression, correlation, and standardized measurements) (Jensen, 1980). Galton was the first researcher to utilize empirically objective devices to measure individual differences in mental abilities (Jensen, 1980). He administered different measures of mental functioning to thousands of individuals as he refined his methods of assessing mental ability. Galton analyzed the scores and applied statistical reasoning to the study of those with high ability. He was the first to identify "general mental ability" in humans (Jensen, 1980).

One of Galton's students, Spearman, was the first to assert that all individual variance in higher order mental abilities is correlated positively. The aforementioned contention supported Galton's belief in a general factor of mental ability (Jensen, 1980). Spearman introduced factor analysis, in part, to ascertain the degree to which a test measures a general factor (Jensen, 1980). Spearman used factor analysis to determine whether the shared variance in a matrix of correlation coefficients results in a single general factor or in several independent more specific factors (Gould, 1996). Spearman believed each test of mental abilities has a single general factor, g, as well as specific factors (s) unique to the test. These beliefs led to the development of the two-factor theory of intelligence. Spearman and many scholars (e.g., Carroll, 1993; Herrnstein &

Murray, 1994; Jensen, 1980; Rushton, 1997) continue to believe scores on intelligence tests are reflected best by g. These theorists consider g to be the most parsimonious method to describe one's intelligence and thus to use when examining mean IQ differences between African-Americans and Caucasian-Americans (Neisser, 1998).

Factor analysis soon became one of the most important techniques in modern multivariate statistics (Gould, 1996; Kamphaus, Petosky, & Morgan, 1997). The technique is useful to reduce a complex set of correlations into fewer dimensions by factoring a matrix of correlation coefficients (Gould, 1981). The variables most highly correlated are combined to form the first principal component by placing an axis through all the points. Other axes, drawn to account for the other variables, are labeled second and third (etc.) order factors.

Relative to intelligence testing, factor analysis has been applied to show positive correlations among different mental tests (Gould, 1996). In that most correlation coefficients in mental tests are positive, factor analysis yields a reasonably strong first principal component (Gould, 1996).

General factor theorists such as Spearman use factor analytic techniques to demonstrate the viability of g as the first factor to emerge when analyzing factor scores for intelligence tests. Other theorists use factor analysis to suggest IQs depend on a number of independent factors, not a large general factor (Gardner, 1983; Spearman, 1923).

Although researchers may disagree about the structure of intelligence, they agree that IQs arise as a function, at least to some degree, from a general factor as well as reflect multidimensional aspects of intellectual functioning (Carroll, 1993; Sattler, 1998;

Urbach, 1974). To reiterate, g is important because it is considered the best way to express one's general mental ability.

Theories of Intelligence

The Cattell-Horn-Carroll theory of intelligence, one of psychology's most recent and comprehensive theories, provides the framework for this study. The theory's historical antecedents can be found in Spearman's two-factor theory of intelligence (Spearman, 1927) and Thurstone's multifactorial theory of intelligence (Thurstone, 1938; Thurstone & Thurstone, 1941). Additionally, it integrates Cattell and Horn's fluid and crystallized theory of intelligence (Horn & Cattell, 1966; Horn & Noll, 1997) and Carroll's Three-Stratum Theory of cognitive abilities (Carroll, 1997, 1993). These theories are described below.

Spearman's g

As noted above, Spearman's theory of intelligence underscores a general factor (g) and one or more specific factors (s). According to Spearman and other general factor theorists, an intelligence test's g loading commonly is most explicative of an individual's attainment on measures of intellectual functioning (Sattler, 1988). Spearman viewed g as general mental energy and that complex or higher order mental activities require the greatest amount of g (Sattler, 1988). The g factor involves mental operations that are generally deductive and associated with the skill, speed, intensity, and amount of an individual's intellectual production (Sattler, 1988).

Spearman identified three major laws of cognitive activities he believed were associated with g.

> The first was the Law of Apprehension, that is, the fact that a person approaches the stimulation he receives from all external and internal sources via the ascending nerves. . . . Next we have the eduction of Relations. Given two

> stimuli, ideas, or impressions, we can immediately discover any relationship existing between them–one is larger, simpler, stronger or whatever than the other. And finally, we have the eduction of Correlates–given two stimuli, joined by a given relation, and a third stimulus, we can produce a fourth stimulus that bears the same relation to the third as the second bears to the first. . . . If Spearman is right, then tests constructed on these principles, that is, using apprehension, eduction of relations and eduction of correlates, should be the best measures of g; that is, correlate best with all other tests. This has been found to be so; the Matrices test. . . has been found to be just about the purest measure of IQ. (Eysenck, 1998, p. 57)

Matrices tests such as the Raven's Progressive Matrices employ Spearman's theory and have been widely used as measures of intelligence (Eysenck, 1998). Matrices tests contain substantial loadings of g and demand conscious and complex mental effort, often evident in analytical, abstract, and hypothesis-testing tasks (Sattler, 1988). Conversely, tests that require less conscious and complex mental effort are low in g (Sattler, 1988). Intelligence tests with lower g emphasize specific factors such as recognition, recall, speed, visual-motor abilities, and motor abilities (Sattler, 1988).

Thurstone's Primary Mental Abilities

Thurstone's (1938) theory of intelligence differs considerably from Spearman's in that Thurstone viewed intelligence as a multidimensional rather than a unitary trait. Thurstone developed the Primary Mental Abilities Test to measure qualities he believed were primary mental abilities: verbal, perceptual speed, inductive reasoning, number, rote memory, deductive reasoning, word fluency, and space or visualization. Thurstone was intent on showing how intelligence could be separated into the noted multiple factors, each of which has equivalent significance (Sattler, 1998). His theory contends that human intelligence is organized systematically with configurations that can be explicated by statistically analyzing the forms of intercorrelations found in a group of tests (Sattler, 1988). Thurstone initially discounted a general factor as a component of

mental functioning. However, because his seven primary factors are moderately correlated, he later came to accept the notion of a second-order factor, g (Sattler, 1988).

Cattell and Horn: Fluid and Crystallized Intelligence

Cattell and Horn (Cattell, 1963; Horn, & Cattell, 1967; Horn & Cattell, 1967) developed a theory of intelligence. Their theory is based on two factors, fluid and crystallized abilities.

> Fluid intelligence refers to essentially nonverbal, relatively culture-free mental efficiency, whereas crystallized intelligence refers to acquired skills and knowledge that are strongly dependent for their development on exposure to culture. Fluid intelligence involves adaptive and new learning capabilities and is related to mental operations and processes, whereas crystallized intelligence involves overlearned and well-established cognitive functions and is related to mental products and achievements. (Sattler, 1992, p. 48)

Fluid intelligence is measured by tasks requiring inductive, deductive, conjunctive, and disjunctive reasoning to understand, analyze, and interpret relationships among stimuli. Crystallized intelligence is measured by tasks requiring acculturation. That is, crystallized intelligence requires familiarity with the salient culture through such qualities as vocabulary and general information. Tests that measure the ability to manipulate information and problem-solving are considered measures of fluid ability whereas tests that require simple recall or recognition of information are considered measures of crystallized abilities (Sattler, 1998).

Carroll's Three-Stratum Theory of Cognitive Abilities

Researchers are making substantial advances each decade in a drive to understand the structure of human intellect. Carroll's (1993) development of a three-stratum theory of intelligence is crucial to these advances. Carroll's book, Human Cognitive Abilities: A Survey of Factor-analytic Studies, summarizes his survey and examination of 460 data

sets, including the majority of important and classic studies of human cognitive abilities (McGrew, 1997). Carroll used exploratory factor analysis to test his belief that human cognitive abilities could be conceptualized hierarchically (McGrew & Woodcock, 2001).

Carroll's work has received highly favorable reviews (Burns, 1994; Eysenck, 1994; and Sternberg, 1994). Currently, there is little objection to his three-stratum theory. The three-Stratum theory is so well received that McGrew noted "simply put, all scholars, test developers, and users of intelligence tests need to become familiar with Carroll's treatise on the factors of human abilities" (McGrew, 1997, p 151). Figure 1-1 and table 1-1 illustrate Carroll's three strata theory.

> The Three-Stratum Theory of cognitive abilities is an expansion and extension of previous theories. It specifies what kinds of individual differences in cognitive abilities exist and how those kinds of individual differences are related to one another. It provides a map of all cognitive abilities known or expected to exist and can be used as a guide to research and practice. It proposes that there are a fairly large number of distinct individual difference in cognitive ability, and that the relationships among them can be derived by classifying them into three different Strata: Stratum I, "narrow" abilities; Stratum II, "broad" abilities; and Stratum III, consisting of a single "general" ability" (Carroll, 1997, p. 122).

> The three-Stratum theory emphasized the multifactorial nature of the domain of cognitive abilities and directs attention to many types of abilities usually ignored in traditional paradigms. It implies that individual profiles of ability are much more complex than previously thought, but at the same time it offers a way of structuring such profiles, by classifying abilities in terms of Strata. Thus, a general factor is close to former conceptions of intelligence, whereas second-Stratum factors summarize abilities in such domains as visual and spatial perception. Nevertheless, some first-Stratum abilities are probably of importance in individual cases, such as the phonetic coding ability that is likely to describe differences between normal and dyslexic readers. (Carroll, 1997, p. 128)

Cattell-Horn-Carroll Theory of Intelligence

The Cattell-Horn-Carroll theory of intelligence is most closely derived from Spearman's theory of g, the fluid and crystallized intelligence theories of Cattell and Horn, and the factor-analytic work of Carroll. McGrew proposed the integrated Carroll

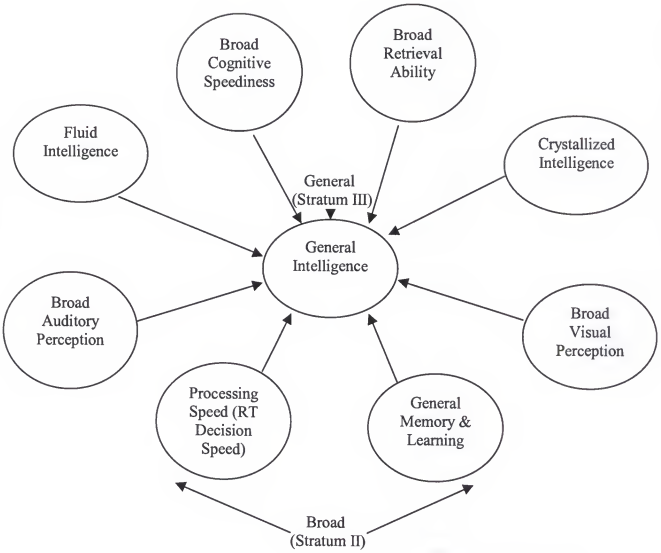

Figure 1-1. Carroll's Strata II and III

Table 1-1

Carroll's Stratum I: Each Narrow Ability is Subsumed Under a Broad Ability

Broad Stratum II Ability	Narrow Stratum I Ability
Fluid Intelligence (Gf)	General Sequential Reasoning (RG) Induction (I) Quantitative Reasoning (RQ) Piagetian Reasoning (RP) Speed of Reasoning (RE?)
Quantitative Knowledge (Gq)	Math Knowledge (KM) Math Achievement (A3)
Crystallized Intelligence (Gc)	Language Development (LD) Lexical Knowledge (VL) Listening Ability (LS) General (verbal) Information (KO) Information about Culture (K2) General Science Information (K1) Geography Achievement (A5) Communication Ability (CM) Oral Production & Fluency (OP) Grammatical Sensitivity (MY) Foreign Language Proficiency (KL) Foreign Language Aptitude (LA)
Reading/Writing (Grw)	Reading Decoding (RD) Reading Comprehension (RC) Verbal (printed) Language Comprehension (V) Cloze Ability (CZ) Spelling Ability (SG) Writing Ability (WA) English Usage Knowledge (EU) Reading Speed (RS)
Short-Term Memory (GSM)	Memory Span (MS) Learning Abilities (l1)

Table 1-1. Continued

Carroll's Stratum I: Each Narrow Ability is Subsumed Under a Broad Ability

Broad Stratum II Ability	Narrow Stratum I Ability
Visual Processing (Gv)	Visualization (VZ) Spatial Relations (SR) Visual Memory (MV) Closure Speed (CS) Flexibility of Closure (CF) Spatial Scanning (SS) Serial Perceptual Integration (PI) Length Estimation (LE) Perceptual Illusions (IL) Perceptual Alternations (PN) Imagery (IM)
Auditory Processing (Ga)	Phonetic Coding (PC) Speech Sound Discrimination (US) Resistance to Auditory Stimulus Distortion (UR) Memory for Sound Patterns (UM) General sound Discrimination (U3) Temporal Tracking (UK) Musical Discrimination & Judgment (U1, U9) Maintaining & Judging Rhythm (U8) Sound-Intensity/Duration Discrimination (U6) Sound Frequency Discrimination (U5) Hearing & Speech Threshold Factors (UA, UT, UU) Absolute Pitch (UP) Sound Localization (UL)
Long-Term Storage & Retrieval (Glr)	Associative Memory (MA) Meaningful Memory (MM) Free Recall Memory (M6) Ideational Fluency (FI) Associational Fluency (FA) Expressional Fluency (FE) Naming Facility (NA) Word Fluency (FW) Figural Fluency (FF) Figural Flexibility (FX) Sensitivity to Problems (SP) Originality/Creativity (FO) Learning Abilities (L1)

Table 1-1. Continued

Broad Stratum II Ability	Narrow Stratum I Ability – continued
Processing Speed (Gs)	Perceptual Speed (P) Rate-of-Test-Taking (R9) Number Facility (N)
Decision/Reaction Time or Speed (Gt)	Simple Reaction Time (R1) Choice Reaction Time (R2) Semantic Procession Speed (R4) Mental Comparison Speed (R7)

and Cattell-Horn model in 1997 (McGrew & Flanagan, 1998). The theory classifies cognitive abilities in three Strata that differ by degree of generality.

Carroll's Stratum I abilities are very similar to the primary factor abilities cited by Horn (1991). Specific abilities within each Stratum positively correlate and thus suggest the different abilities in each Stratum do not reflect completely independent traits (Carroll, 1993; Flanagan & Ortiz, 2001).

> Carroll identifies 69 specific, or narrow, abilities and conceptualized them as Stratum I abilities. These narrow abilities are grouped into broad categories of cognitive ability (Stratum II), which he labeled Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, Broad Auditory Perception, Broad Retrieval Ability, Broad Cognitive Speediness, and Processing Speed. At the apex of his model (Stratum III), Carroll identified a general factor which he referred to as General Intelligence, or "g." (McGrew & Woodcock, 2001, p. 11)

Extensive factor analytic, neurological, developmental, and heritability evidence (Flanagan & Ortiz, 2001) supports the Cattell-Horn-Carroll theory of intelligence. In addition, recent research suggests the theory provides equal explanatory power across gender and ethnicity (Carroll, 1993; Gustafsson & Balke, 1993; Keith, 1997 & 1999). "In general, the CHC theory is based on a more thorough network of validity evidence than other contemporary multidimensional models of intelligence" (Flanagan & Ortiz, 2001, p. 8). The WJ-III is the only intelligence test based extensively on CHC theory (Keith, Kranzler, & Flanagan, 2001) and, as such, will be the instrument under study in this research.

Purpose of the Study

This study investigates possible IQ differences between African-Americans and Caucasian-Americans for all combined ages on the Woodcock Johnson-III: Tests of Cognitive Abilities in view of the recently developed Cattell-Horn-Carroll theory of

intelligence. In addition, the factor structure and IQ-achievement correlations for the WJ-III will be investigated for the groups. These two groups are studied because they are two of the largest racial groups in the United States. African-Americans constitute roughly 13% of the U.S. population (U.S. Census, 2000). Prior research indicates the mean IQ of African-Americans is more than 15 points below that for Caucasian-Americans on tests of pure g (Jaynes & Williams, 1989; Jensen, 1980). The term Spearman's hypothesis was coined to identify this theory, which postulates mean IQ differences among subgroups occur as a function of intelligence tests' g loadings (Jensen, 1998). The term Spearman-Jensen hypothesis will be used in this study to reflect the theory that mean IQ differences among subgroups occur as a function of intelligence tests' g loadings.

Jensen was one of the most influential researchers to suggest African-Americans tend to score lower than Caucasian-Americans on g loaded tests (Stratum III) than on tests of narrow (Stratum I) and broad (Stratum II) abilities. Jensen noted "[m]y perusal of all the available evidence leads me to the hypothesis that it is the item's g loading, rather than the verbal-nonverbal distinction per se, that is most closely related to the degree of white-black discrimination of the item" (Jensen, 1980, p. 529). Jensen indicated IQ differences between African-Americans and Caucasian-Americans on published mental tests are most closely related to the g component in score variance and do not result from the tests' factor structure, cultural loading, or test bias (Jensen,1980). That is, variation in mean differences between the two groups cannot be explicated based on the tests' item content or any formal or superficial characteristics of the tests (Jensen, 1998). Intelligence tests in common use have the same reliability and validity for native, English-speaking African-Americans as they have for Caucasian-Americans (Jensen,

1998). The degree of the test's g loading predicts the magnitude of the standardized mean subgroup difference (Jensen, 1998).

Two additional factors, aside from g, also reveal differences between the two groups. On average, African-Americans obtain higher scores than Caucasian-Americans on tests of short-term memory. On the other hand, Caucasian-Americans, on the average exceed African-Americans on tests of spatial visualization (Jensen, 1998). "The effects of these factors, however, show up only on tests that involve these factors, whereas the g factor enters into the W-B differences on every kind of cognitive test" (Jensen, 1998, p. 352).

The magnitude of differences between African-Americans and Caucasian-Americans is expected to be smaller than the traditional 15 points on tests based on, or consistent with, the Cattell-Horn-Carroll theory of intelligence. In addition, based on Jensen's (1998; 1980) work, it is likely the factor structure and IQ-achievement correlations will not differ for the two groups. Support for the smaller mean difference hypothesis is found below.

The WJ-III, as a CHC theoretical measure, is comprised of specific and broad abilities. Specificity refers to the proportion of a test's true-score variance that is unaccounted for by a common factor such a g (Jensen, 1998). On most intelligence tests, approximately 50% of the variance of each subtest is specific to that subtest. As such, its source of variance is partly comprised by g and is partly separate of g (Jensen, 1998). IQ differences between African-Americans and Caucasians should be smaller than 15 points on intelligence tests comprised of specific (i.e., Stratum I), or broad (i.e., Stratum II) abilities (tests consistent with CHC theory such as the WJ-III). Again, the aforementioned thesis has extensive support based on the Spearman-Jensen hypothesis

(Jensen, 1998). To reiterate, in light of the specific and broad factors on tests based on CHC theory, their g loadings are smaller.

Further support for reduced mean IQ differences among racial/ethnic groups on the WJ-III is evident in data from the Kaufman-Assessment Battery for Children (K-ABC), a multi-factor intelligence test that has lower g loadings than many other measures of intelligence (Bracken, 1985). Data from the K-ABC's standardization sample indicate African-Americans scored approximately one-half standard deviation below Caucasian-Americans on the K-ABC (Kaufman & Kaufman, 1983). The K-ABC does not utilize a hierarchical theory of intelligence and instead centrally assesses multiple specific abilities (Kaufman & Kaufman, 1983).

The hierarchical structure of the WJ-III includes multiple specific and broad abilities that suggests it has relatively lower g loadings than some other intelligence tests (e.g., the Wechsler Intelligence Scale for Children – Third Edition, the Differential Abilities Scale, and the Standford-Binet Fourth Edition). Nonetheless, the test is considered a robust measure of g (Flanagan & Ortiz, 2001).

Data regarding the factor structure of the WJ-III are reported for African-Americans and Caucasian-Americans. The test authors report a root mean square error of approximation (RMSEA) fit statistic of .039 for the two groups (McGrew & Woodcock, 2001), which suggests the WJ-III measures the same constructs for Caucasian and non-Caucasians in the standardization sample.

Data relative to mean IQ differences between African-Americans and Caucasian-Americans and IQ-achievement correlations are not reported, by group, for African-Americans and Caucasian-Americans on the WJ-III. IQ-achievement correlations will be investigated to determine whether correlations will not differ between the GIA and the

Broad Reading, the GIA and the Broad Math, and the GIA and Broad Written Language factors for the two groups. Given the Spearman-Jensen hypothesis, IQ-achievement correlations will likely not differ for African-Americans and Caucasian-Americans on the WJ-III. Additionally, in light of the WJ-III's specific and broad abilities (Carroll's Strata I and II), the mean IQs of African-Americans and Caucasians are likely to differ, but by fewer than 15 points.

CHAPTER 2
REVIEW OF THE LITERATURE

The Development of Intelligence

Scholars have yet to reach consensus as to the best definition of intelligence. Lack of consensus has led to difficulty understanding intelligence as a unified construct (Valencia & Suzuki, 2001). Nonetheless, some agreement is evident given the generally accepted view that intellectual development is a function of nature and nurture (Gould, 1996; Plomin, 1988; Sattler, 1992). Both genetic and environmental variables and the interaction between them impact the development of intelligence (Styles, 1999). Additionally, the progression of intellectual development can be viewed as either continuous or discontinuous. When considered continuous, development is connected and smooth. When considered discontinuous, it is interrupted and occurs in spurts.

The psychometric and cognitive-developmental perspectives provide the two theoretical frameworks most often used to understand the development of intelligence (Elkind, 1975). From a psychometric perspective, the development of intelligence is considered continuous. Conversely, from a cognitive-developmental perspective, the development of intelligence is viewed as discontinuous (Epstein, 1974a & b).

To a degree, the psychometric and cognitive-developmental perspectives are complementary because both support the fundamental adaptive role of intelligence and changes are seen as moving in the direction of greater complexity as one enters early adulthood. Intelligence develops on a continuum of increasing capacity (Styles, 1999). However, from a psychometric perspective intelligence is considered generally stable

throughout the life-span (understanding that IQs generally decrease in the elderly), but from a cognitive-developmental perspective, stability of intelligence does not occur until around the age of 15 and beyond (Epstein, 1974a & b).

Styles (1999) indicated children evidence several intellectual growth spurts that occur at different ages, suggesting the spurts are best explained by maturational changes primarily due to nature as opposed to environmental changes that are primarily due to nurture (Andrich & Styles, 1994; Styles, 1999). As Styles noted, "[T]here is not reason that, for example, educational opportunities would directly cause a growth spurt; if it were so, all children would spurt at the same time and if this were so, the pattern of variance would not occur-the variance would remain linear and parallel to the horizontal axis" (1999, p. 31).

Proponents of psychometric theory suggest the development of intelligence can be understood best by using a quantitative perspective of assigning individual scores. The cognitive-developmental theory of intellectual development asserts children develop in stages along a continuum and it is their qualitatively different reasoning abilities that indicate in which stage they operate. Over the decades, psychometric theory became the most prevalent method of measuring intelligence.

Pros and Cons of Intelligence Testing

The first practical intelligence test was developed in 1905 by Binet and Simon as a means of objectively measuring intelligence and diagnosing degrees of mental retardation (Sattler, 1988). Despite its long history, a great deal of ambiguity exists as to appropriate uses of intelligence tests. The ambiguity is associated with the awareness that intelligence is a quality and not an entity, and that, to some degree, the tests measure examinees' prior learning (Wesman, 1968). Additionally, intelligence is a hypothetical

construct that is inferred rather than directly observed (Reynolds, Lowe, & Saenz, 1999). That is, to some degree intelligence is a subjectively determined psychological construct. The aforementioned ambiguity can lead to misuses of intelligence tests and misapplication of test results.

Inappropriate use of intelligence tests can result in the under-utilization of children's potential. For example, children may be labeled improperly and placed in programs for students with educational deficits, denied placement in programs for gifted students, and be subject to reduced educational expectations. Restrictions in educational placement may result in reduced opportunities for minority students to graduate from high school with regular diplomas (Valencia & Suzuki, 2001).

Appropriate intelligence testing aids in diagnosis of handicapping conditions. Intelligence testing helps evaluate programs, reveal inequalities, and provides an objective standard. IQs are helpful in ascertaining present and future functioning. Additionally, IQs assist in the identification of the academic potential of students. Significantly, intelligence test scores can be a great equalizer because the data are able reduce teacher prejudice by using statistically valid standardized tests to ascertain high ability among minority children who may have otherwise been unrecognized. (For a more extensive presentation on the pros and cons of intelligence testing see Sattler, 1988, p. 78.)

The benefits of intelligence testing notwithstanding, test users need to be aware of the influence of intelligence test scores on students' educational placement. Additionally, tests users need information about how specific intelligence test differentially impact minority groups. As DeLeon (1990) and Messick (1995) suggested, for test scores to be construed as fair and valid, they need to be interpreted in light of their statistical validity

and the consequences of the student's performance within the context of culture, language, home, and community environments.

The Cultural Influence on IQ

Learning influences intelligence and thus performance on intelligence tests. As a result, the environment and culture of the examinee that foster or hamper learning becomes important. Moreover, the influence of culture on test scores is important because cultural bias is cited as one major reason why African-Americans earn lower IQs than Caucasians. Of course, culture pertains to more than region, race, ethnicity, or language. Inferring equality of culture based simply on region, race, ethnicity, and language is untenable (Frisby, 1998).

While all tests are influenced by culture, they may not be culturally biased (Sattler, 2001). "Intelligence tests are not tests of intelligence in some abstract, culture-free way. They are measures of the ability to function intellectually by virtue of knowledge and skills in the culture of which they are a sample" (Scarr, 1978 p. 339). Attempts to develop intelligence tests entirely absent the impact of cultural experiences and learning that accrues from these experiences are unlikely (Sattler, 1988). Whether the test is culturally loaded or culturally biased is the important distinction (Jensen, 1974).

Culturally loaded tests require knowledge about specific information important to a particular culture. This knowledge includes awareness of the culture's communication patterns, including verbal and nonverbal representations of the language.

Importantly, a test is considered culturally biased when it measures different abilities for various racial/ethnic groups, when there is a significant difference between its predictive ability for the groups, and when test results are significantly affected by the

differential experience of the groups (Sattler, 1988). Cultural loading is a necessary but insufficient condition for an intelligence test to be considered culturally biased. That is, a culturally loaded test is not necessarily culturally biased. However, tests that are culture loaded or saturated should be analyzed to determine whether the tests measure different abilities for different racial/ethnic groups, differentially predict subgroup performance, and are significantly affected by the different experiences among those who comprise the subgroups.

Statistical analyses of intelligence testing indicate most individual intelligence tests are not culturally biased (Sattler, 1988). However, differences in their cultural loading exist. (Sattler, 1988). Tests that are highly culturally loaded utilize stimuli specific to knowledge or experience associated with a given culture.

In contrast, tests with reduced cultural loading such as the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998) and the Raven's Progressive Matrices are developed to measure problem-solving by utilizing spatial and figural content. These types of tests assess abilities based on experiences that are generally similar to and congruent across ethnic and racial groups and are considered to contain culturally reduced content (Sattler, 2001). The key phrase in the previous sentence is "culturally reduced." Even matrices' tests, such as the Raven's, are not free of cultural influences. Despite their culturally reduced ranking, intelligence tests that emphasize problems involving spatial and figural content tend to be robust measures of g.

Case Law, Cultural Bias, and Intelligence Testing

In Larry P. v. Riles et al. (495 F. supp. 926, N.D. CA. 1979; 793 F. 2d 969, 9th Cir. 1984) a federal court considered intelligence tests culturally biased against minorities to such a degree that the Court ruled that standardized intelligence tests could not be used

to make special educational decisions involving African-American children in California (Opton, 1979; Sattler, 1988). In opposition to the Larry P. decision, in a case from Illinois (Parents in Action on Special Education v. Joseph P. Hannon - PASE, 1980) a federal court found intelligence tests were not biased against cultural and ethnic minorities (Reynolds, et al., 1999; Sattler, 1988). The Larry P. decision later was overturned by a federal appeals court, making case law generally congruous with PASE (Reynolds et al., 1999). Nonetheless, as a result of the Larry P. case, in California the judge's ban remained in force as of September 2000 preventing the use of intelligence tests with children who are being considered, or who are in programs, for the educable mentally handicapped (Sattler, 2001).

Writing about Larry P., Hilliard (1992) emphasized that the judge in the case had concerns about the efficacy of instruction in special education classrooms. Moreover, the judge expressed profound dismay with the general philosophy of education that supported professional practices leading to such inequities as the disproportionate placement of African-American children in classes for the educable mentally handicapped. The judge hoped that his treatise on the use of intelligence tests would be a way to stimulate researchers, professional educators, and psychologists to tackle these fundamental problems with respect to social consequences of testing, rather than merely focusing on the problems of statistical test bias and validity (Hilliard, 1992).

Special Education Eligibility and Intelligence Testing

Several researchers support the assertion that reliance on standardized instruments in the psychological evaluations of students has caused a large number of students to be inappropriately placed in special education programs because of their cultural and linguistic differences (DeLeon, 1990; Finlan, 1994 & 1992; Ysseldyke, Algozzine, &

McGue, 1995). Learning disabilities and mental handicaps are two special education categories considerably impacted by scores from intelligence tests (Valencia & Suzuki, 2001).

With respect to special education classification, researchers in favor of intelligence testing note intelligence testing is only one part of the overall process. As Lambert (1981) indicated, "[I]t is failure in school, rather than tests scores, that initiates action for special education consideration" (p. 940). Moreover, some suggest the disproportionate number of minorities in special education programs is due to the fact minorities are referred much more frequently for special education testing (Reynolds, et al. 1999). Nonetheless, ". . . tests are ubiquitous in psychoeducational assessment and often carry significant implications with respect to questions regarding diagnosis and intervention" (Ortiz, 2000, p.1322).

With the passage of Public Law 94-142, the Education for All Handicapped Children Act, the use of intelligence test in schools became more prominent (Finlan, 1994; & 1992). The law was reauthorized in 1997 as Public Law 101-457, Individuals with Disabilities Education Act – IDEA (IDEA, 1997).

As part of IDEA, a student with academic difficulties is identified as having a learning disability when he or she has an IQ in the average range or higher but whose reading, writing, or arithmetic is well below the expected levels given the obtained IQ. Conversely, a student who evidences academic difficulties but commensurate intellectual ability is not considered learning disabled (IDEA, 1997). Most states use some form of intelligence test score when determinations are made as to a student's eligibility for learning disability services (Frankenberger & Fronzaglio, 1991).

In addition, intelligence tests are used when deciding whether students are eligible for services based on a mental handicap. Students with IQs substantially below the mean, and who also evidenced academic deficits and problems in adaptive functioning are considered mentally handicapped and therefore eligible for services in special education classes (IDEA, 1997).

Of the many reasons for the continued use of IQs in education, two are most salient: First, when the federal government recognized learning disabilities and mental handicaps as educational handicapping conditions, it also provided additional funding to states to assist in the education of students who are in these categories. School districts receive federal funding for students in the district who are enrolled in special education programs (Finlan, 1994; 1992).

Second, IDEA requires students enrolled in special education programs to participate in state and district-wide group standardized assessments of academic achievement. Nonetheless, scores for students in special education programs often are disaggregated from those from the general student population for reporting purposes (U.S. Department of Education, Office for Civil Rights, 2000). Schools that are able to disaggregate a greater number of scores from the general student population tend to obtain higher overall group scores on the state-wide achievement tests and may be considered higher performing schools.

For approximately 10 years California was not allowed to use intelligence tests to determine African-American students' eligibility for special education program. During the noted period, the proportion of African-American students placed in mentally handicapped and developmentally delayed programs decreased, but the proportion placed in programs for students with learning disabilities increased (Morison, White, & Feuer,

1996). Thus, the use of intelligence tests impacts the proportions of African-Americans placed in specific special education programs.

Clearly, there are administrative and diagnostic reasons for the extensive use of intelligence tests in schools (Aaron, 1997; Finlan, 1994, 1992; Ysseldyke, Algozzine, & McGue, 1995). These administrative and diagnostic reasons, in tandem with Child Find legislature (the requirement for states to locate potentially disabled children), conceivably led to the upsurge in enrollment of students in special education programs across the United States (Finlan, 1994). Over the last 10 years, there was an approximately 35 % upsurge in the numbers of children served under IDEA (Donovan & Cross, 2002). All of the aforementioned establish, at least in part, reasons why intelligence testing continues to be widely valued in education.

Overrepresentation of Minorities in Special Education

Available data suggest minorities are overrepresented in some special education programs. Overrepresentation is not operationally defined and seems to refer to any percentage difference in special education participation and presence in the general population by race/ethnicity. Perhaps it would be helpful for experts to operationally define overrepresentation. Although determinations as to overrepresentation are arbitrarily assigned, a difference of 20% or more is certainly notable. Such a difference likely does not occur exclusively as a function of chance.

The 1998-1999 school year was the first year the federal government required states to report on the incidence of minorities in special education programs. African-Americans comprise approximately 15% of the nation's population, but roughly 34% of students in the mentally handicapped program. The difference is about 19% and for the purposes of this study 20% will be considered the cut-score to define disproportionate

representation in the educable mentally handicap category. The state of Florida uses a similar procedure. The term disproportionate representation will be used in this study to indicate participation in special education that differs from the subgroups' presence in the resident population by 20% or greater. As a consequence, overrepresentation is evident in states and school districts when African-Americans comprise a proportion of 20% or greater of students in mentally handicapped programs than in the general population. In the context of this study, an operational definition of disproportionate representation is not terribly critical. Rather, disproportionate representation is highlighted in reference to the consequential validity or social consequences of IQ. The greater the mean difference among subgroups, the greater the negative social consequences.

Table 2-1 presents data from U.S. Department of Education's Twenty-second Annual Report to Congress on the Implementation of the Individuals With Disabilities Education Act (2000) relative to the incidence of mental handicaps classification by racial/ethnic group across the nation. African-American (non-Hispanic) students total 15% of the general population for ages 6 through 21, compared with 20% of the special education population among all disabilities. African-American students' representation in the mental retardation category was more than twice their national population estimates (15% v. 34%). Representation of Hispanic students in special education (13%) was generally similar to the percentages in the general population (14%). Native American students represent 1% of the general population and 1.3% of special education students. Overall, white (non-Hispanic) students made up a slightly smaller percentage (64%) of the special education students than the general population (66%).

Comparisons of the racial/ethnic distribution of students in special education with the general student population reveal Asian and Caucasian students were represented at a lower rate than their presence in the resident population. Native American and African-American students were represented in special education at a higher rate than their presence in the resident population. Hispanic students generally were represented in special education at a rate comparable to their proportion of the U. S. population (U.S. Department of Education, Twenty-second Annual Report to Congress on the Implementation of the Individuals With Disabilities Education Act, Office of Special Education Programs, 2000).

Figures on the disproportionate representation of minorities in special education categories have been criticized for several reasons. For example, the data for some minority groups frequently vary based on the groups reporting or interpreting the data (Artiles & Trent, 1994). Differing statistical analyses may be used in different studies (Valencia & Suzuki, 2001). Additionally, as Reschly (1981) noted, "Analyses of overrepresentation have largely ignored the variables of gender and poverty as well as the other steps in the referral-placement process" (p. 1095). A correlation is apparent between SES and placement in LD and mentally handicapped programs (Brosman, 1983).

Despite the problems associated with understanding disproportionate representation, the overrepresentation of African-American students in special education categories is problematic because these students frequently operate in restrictive educational placements that may not be most conducive to their learning (Valencia &

Table 2-1

Percentage of Students Ages 6 Through 21 Served by Disability and Race/Ethnicity in the 1998-99 School Year

Disability	NA	API	AA	H	W
Autism	.7	4.7	20.9	9.4	64.4
Deaf-Blindness	1.8	11.3	11.5	12.1	63.3
Developmental Delay	.5	1.1	33.7	4.0	60.8
Emotional Disturbance	1.1	1.0	26.4	9.8	61.6
Hearing Impairments	1.4	4.6	16.8	16.3	66.0
Mental Handicaps	1.1	1.7	34.3	8.9	54.1
Multiple Disabilities	1.4	2.3	19.3	10.9	66.1
Orthopedic Impairments	.8	3.0	14.6	14.4	67.2
Other health Impairments	1.0	1.3	14.1	7.8	75.8
Specific Learning Disabilities	1.4	1.4	18.3	15.8	63.0
Speech and Language Impairments	1.2	2.4	16.5	11.6	68.3
Traumatic Brain Injury	1.6	2.3	15.9	10.0	70.2
Visual Impairments	1.3	3.0	14.8	11.4	69.5
All Special Education Disabilities	1.3	1.7	20.2	13.2	63.6
Resident Population	1.0	3.8	14.8	14.2	66.2

Key: NA = Native American; API= Asian/ Pacific Islander; AA = African-American (non- Hispanic); H = Hispanic; W = White (non-Hispanic)

Source: U.S. Department of Education, Twenty-second Annual Report to Congress on the Implementation of the Individuals With Disabilities Education Act. (2000). Office of Special Education Programs, Data Analysis System (DANS).

Suzuki, 2001). Disproportionate representation of African-Americans in special education programs essentially results in the segregation of students, which is in direct opposition to current American values and federal case law.

Among several other reasons, states differ with respect to the prevalence of students enrolled in special education programs because psychologists use different measurement devices when evaluating students. Additionally, within the context of federal law, each state decides what specific criteria are important when diagnosing learning disabilities and mental handicaps and how it wishes to administer its educational programs for students diagnosed with these conditions. For example, a student could be diagnosed as learning disabled based on an IQ of 80 (the 9th percentile) or above in one state and with an IQ of 85 (the 16th percentile) or above in another state (Finlan, 1994; 1992). Moreover, an IQ of 75 (the 5th percentile) or below (coupled with deficient adaptive behavior skills) could result in placement in a mentally handicapped program in one state and whereas an IQ of 69 below is needed in another. Thus, a relatively small difference in IQ can have a large impact on students' educational placement.

To reduce disproportionate representation as a result of inadvertent bias, tests users need to know which intelligence tests best represent and most reliably and fairly reflect minority group scores. The selection and administration of intelligence tests and the interpretation of their scores should be based on substantial research and test fairness information, otherwise decision-making as a function of the resultant data may be biased and materially untenable (Sandoval, 1998).

Test Bias

Bias in mental testing is an important issue to consider when discussing mean IQ differences. Bias in testing essentially concerns the presence of construct irrelevant components and construct underpresentation in tests that produce systematically lower or higher scores for subgroups of test takers (American Educational Research Association, et al., 1999). Relevant subgroups are characterized on the basis of race, ethnicity, first language, or gender (Scheuneman & Oakland, 1998). Scholars often describe two forms of test bias or error: random and systematic error. Random error occurs on all tests to some degree and is due to such conditions as examinee fatigue and measurement error. Random errors also occur as a function of test session behavior. For example, examinee attentiveness, nonavoidance of task, and cooperative mood were found to be significantly related to student performance on individually administered measures of intelligence and achievement (Glutting & Oakland, 1993). Examinees who demonstrate low levels of the noted qualities tend to score lower on intelligence and achievement tests (Scheuneman & Oakland, 1998).

Systematic errors reflect problems in the development and/or norming of intelligence tests such as inappropriate sampling of test content or unclear test instructions. Test content problems such as construct underrepresentation refers to a rather narrow sampling of the dimensions of interest. Construct-irrelevant variance occurs when an irrelevant task characteristic differentially impacts subgroups. It refers to overly broad and immaterial items sampling of the facets of the construct that may increase the difficulty or easiness of the task for individuals or groups (American Educational Research Association, et al., 1999; Messick, 1995).

Test developers attempt to minimize both forms of error (Frisby, 1999; Sattler, 2001). Attempts to attenuate bias and error in the development and use of intelligence tests are necessary in light of the fact these tests frequently are used and significantly influence diagnosis, placement, and intervention with students experiencing school problems (Ortiz, 2000). Nonetheless, all intelligence tests contain some degree of error and thus never are completely reliable. Tests biased in favor of the majority will substantially impact mean score differences among subgroups (American Educational Research Association, et al., 1999; Messick, 1995; Reynolds et al., 1999; Sattler, 2001).

In fact, when using grouped data, intelligence tests tend to underestimate the academic performance of Caucasians and overestimate the academic performance of African-Americans (Braden, 1999). Given the aforementioned, some might suggest when intelligence tests are used to predict academic achievement they are biased in favor of African-Americans. Proportionately, African-Americans students are much more likely to be negatively impacted by test score use. Therefore, these tests are subject to predictive bias, which is the systematic under- or over-prediction of criterion performance for persons belonging to groups differentiated by characteristics not relevant to criterion performance (American Educational Research Association, et al., 1999). Tests used in education that contain predictive bias may not offer sufficient utility to support their continued use.

Nonetheless, the purpose of this study is not to suggest the WJ-III or any of the well-standardized and popular intelligence tests are biased against persons from some minority groups. To reemphasize, this study is not designed to test or measure bias on the WJ-III. The test authors reported factor invariant data that suggest the instrument is not biased against relevant subgroups in reference to construct validity. However, when test

users are unaware of the mean IQ differences for relevant subgroups on intelligence tests in common use, the *testing process* itself may lack sufficient social validity, appear biased, and may be detrimental to lower scoring groups. One goal of this study is to provide knowledge of mean score differences so as to allow practitioners a degree of influence in decreasing the consequential impact or increasing the social validity of test scores. As Jensen (1998) noted:

> For groups, the most important consequence of a group difference in means is of a statistical nature. This may have far-reaching consequences for society, depending on the variables that are correlated with the characteristic on which the groups differ, on average, and how much society values them. In this statistical sense, the consequences of population differences in IQ (irrespective of cause) are of greater importance, because of all the important correlates of IQ, than are most other measurable characteristics that show comparable population differences. (p. 354)

Researchers who oppose the use of intelligence tests view validity from a social/cultural framework, while researchers who support the use of intelligence tests view validity using a predominantly statistical framework. Messick's (1995) work integrated the two frameworks.

Recent Concepts of Test Validity

Traditional concepts of validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; Geisinger, 1998; Reynolds et al., 1999) considered content, construct, and criterion as three major and different aspects of validity. Recently, many scholars have come to consider these concepts somewhat fragmented and incomplete (American Educational Research Association, et al., 1999; Messick, 1995). Current scholarship describes validity in reference to psychometric and statistical properties as well as a social concept. Validity as a psychometric and statistical concept reflects norming procedures, reliability,

content validation, criterion-related validation, and construct validation (Geisinger, 1998). Validity as a social concept considers notions as to whether intelligence tests measure past achievement or ability for future achievement and the resulting social consequences of score use.

Messick (1995) recognized the importance of validity, reliability, comparability, and fairness and believed these four concepts also embody social values that are meaningful (even aside from assessment) whenever appraisals and conclusions are reached. He supported the predominant view that validity is not a property of the test or assessment as such but of the meaning derived from test scores.

> Indeed, validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Therefore, it is fundamental that score validation is an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use. (Messick, 1995, p 742)

Lack of understanding as to the social consequences of intelligence test scores can lead to bias in mental testing. According to DeLeon (1990), assessment practices based on the philosophies of examiners is the least discussed issue in the literature. For example, although tradition plays a part in test selection, examiners' philosophical orientation also determines which intelligence test examiners chose to administer. Determinations about the manner in which evaluations should be conducted and the types of data that are most important can ultimately lead to appropriate (nonbiased) as well as

inappropriate (biased) evaluations of minority children without any intentional biases on examiners' part (DeLeon, 1990).

Social Validity

Examiners make decisions as to whether culture-reduced, culture-loaded, high g, or low g tests are administered. Examiners also determine whether a verbal or nonverbal test should be administered. Consequently, it is important to provide as much data as readily available on the fairness and social consequences of intelligence test scores to assist psychologists make decisions concerning which are the most reliable, valid, and fair intelligence tests to administer. As Oakland and Laosa (1976) noted, "test misuse generally occurs when examiners do not apply good judgment… governing the proper selection and administration of tests" (p. 17).

The importance of considering the social consequences of intelligence testing, both intended and unintended, when intelligence tests produce substantial differences in mean IQs among racial/ethnic subgroups, also is highlighted (The standards for Educational and Psychological Testing: (heretofore The standards) standard 13.1; American Educational Research Association, et al., 1999; Messick, 1995).

> Evidence about the intended and unintended consequences of test use can provide important information about the validity of the inferences to be drawn from the test results, or it can raise concerns about an inappropriate use of a test where the inferences may be valid for other uses. For instance, significant differences in placement test scores based on race, gender, or national origin may trigger a further inquiry about the test and how it is being used to make placement decisions. The validity of the test scores would be called into question if the test scores are substantially affected by irrelevant factors that are not related to the academic knowledge and skills that the test is supposed to measure. (U.S. Department of Education, Office for Civil Rights, 2000, p. 35)

Psychological assessment of school age children often depends heavily on the use of standardized intelligence tests. Attempts to consider the social and value implications

of IQ meaning and use require test users know the mean IQ differences for various racial/ethnic groups and the standard deviations of their distributions. As noted by OCR,

> When tests are used as part of decision-making that has high-stakes consequences for students, evidence of mean score differences between relevant subgroups should be examined, where feasible. When mean differences are found between subgroups, investigations should be undertaken to determine that such differences are not attributable to construct underrepresentation or construct irrelevant error. *Evidence about differences in mean scores and the significance of the validity errors should also be considered when deciding which test to use.* (U.S. Department of Education, Office for Civil Rights, 2000, p. 45; emphasis added)

Knowledge of mean IQ differences allows test users to determine whether specific intelligence tests may impact racial/ethnic groups differentially.

It is important for test publishers and researchers to furnish test users with as much information as possible about mean score differences to help them make knowledgeable and fair decisions to effectively utilize intelligence test scores when evaluating children (American Educational Research Association et al., 1999). According to standard 7.11 (American Educational Research Association, et al., 1999, p. 83), "[W]hen a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, *evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use* (emphasis added)." Test scores will likely continue to be of substantial importance in high-stakes decision making in education (Scheuneman & Oakland, 1998). Therefore, the use of each intelligence test must be guided by substantial research, including research on subgroup differences. The results that address hypotheses that guide this study have the potential of adding to the research database in this area. The following hypotheses will be tested:

Statement of Hypotheses

1. The factor structure of the WJ-III will not differ appreciably for African-Americans and Caucasian-Americans.

2. Mean scores on the WJ-III General Intellectual Ability factor, Stratum III, will be higher for Caucasian-Americans than African-Americans.

3a. Mean scores on the WJ-III test of Verbal Comprehension will be higher for Caucasian-Americans than for African-Americans.

3b. Mean scores on the WJ-III Visual-Auditory Learning will be higher for Caucasian-Americans than for African-Americans.

3c. Mean scores on the WJ-III Spatial Relations will be higher for Caucasian-Americans than for African-Americans.

3d. Mean scores on the WJ-III Sound Blending will be higher for Caucasian-Americans than for African-Americans.

3e. Mean scores on the WJ-III Concept Formation will be higher for Caucasian-Americans than for African-Americans.

3f. Mean scores on the WJ-III Visual Matching will be higher for Caucasian-Americans than for African-Americans.

3g. Mean scores on the WJ-III Numbers Reversed will be higher for Caucasian-Americans than for African-Americans.

4. Mean score difference on the WJ-III General Intellectual Ability factor between Caucasian-Americans and African-Americans will be less than 15 points.

5a. Mean differences between African-Americans and Caucasian-Americans will be less on Verbal Comprehension than on general intelligence.

5b. Mean differences between African-Americans and Caucasian-Americans will be less on Visual-Auditory Learning than general intelligence.

5c. Mean differences between African-Americans and Caucasian-Americans will be less on Spatial Relations than on general intelligence.

5d. Mean differences between African-Americans and Caucasian-Americans will be less on Sound Blending than on general intelligence.

5e. Mean differences between African-Americans and Caucasian-Americans will be less on Concept Formation than on general intelligence.

5f. Mean differences between African-Americans and Caucasian-Americans will be less on Visual Matching than on general intelligence.

5g. Mean differences between African-Americans and Caucasian-Americans will be less on Numbers Reversed than on general intelligence.

6a. General intelligence and Broad Reading will correlate significantly for African-Americans and Caucasian-Americans.

6b. Correlations between general intelligence and Broad Reading will not differ for African-Americans and Caucasian-Americans.

6c. General intelligence and Letter-Word Identification will correlate significantly for African-Americans and Caucasian-Americans.

6d. Correlations between general intelligence and Letter-Word Identification will not differ for African-Americans and Caucasian-Americans.

6e. General intelligence and Reading Fluency will correlate significantly for African-Americans and Caucasian-Americans.

6f. Correlations between general intelligence and Reading Fluency will not differ for African-Americans and Caucasian-Americans.

6g. General intelligence and Passage Comprehension will correlate significantly for African-Americans and Caucasian-Americans.

6h. Correlations between general intelligence and Passage Comprehension will not differ for African-Americans and Caucasian-Americans.

7a. General intelligence and Broad Math will correlate significantly for African-Americans and Caucasian-Americans.

7b. Correlations between general intelligence and Broad Math will not differ for African-Americans and Caucasian-Americans.

7c. General intelligence and Calculation will correlate significantly for African-Americans and Caucasian-Americans.

7d. Correlations between general intelligence and Calculation will not differ for African-Americans and Caucasian-Americans.

7e. General intelligence and Math Fluency will correlate significantly for African-Americans and Caucasian-Americans.

7f. Correlations between general intelligence and Math Fluency will not differ for African-Americans and Caucasian-Americans.

7g. General intelligence and Applied Problems will correlate significantly for African-Americans and Caucasian-Americans.

7h. Correlations between general intelligence and Applied Problems will not differ for African-Americans and Caucasian-Americans.

8a. General intelligence and Broad Written Language will correlate significantly for African-Americans and Caucasian-Americans.

8b. Correlations between general intelligence and Broad Written Language will not differ for African-Americans and Caucasian-Americans.

8c. General intelligence and Spelling will correlate significantly for African-Americans and Caucasian-Americans.

8d. Correlations between general intelligence and Spelling will not differ for African-Americans and Caucasian-Americans.

8e. General intelligence and Writing Fluency will correlate significantly for African-Americans and Caucasian-Americans.

8f. Correlations between general intelligence and Writing Fluency will not differ for African-Americans and Caucasian-Americans.

8g. General intelligence and Writing Samples will correlate significantly for African-Americans and Caucasian-Americans.

8h. Correlations between general intelligence and Writing Samples will not differ for African-Americans and Caucasian-Americans.

The expectation of reduced mean IQ differences between African-Americans and Caucasian-Americans on the WJ-III is based on the Spearman-Jensen hypothesis and CHC theory. As previously discussed, the Spearman-Jensen hypothesis suggests IQ differences between African-Americans and Caucasian-Americans on mental tests are thought to be related most closely to the g component in score variance, not to cultural loading, specific factors, or test bias (Jensen, 1998; 1980).

CHAPTER 3
METHODS

Participants

The data used in this study include 1,975 Caucasian-Americans and 401 African-Americans who participated in the standardization of the WJ-III. Participants were selected from more than 100 geographically diverse communities in the north, south, west and midwest regions of the United States. An additional 775 participants were administered combinations of the 42 WJ-III tests concurrently with other tests' batteries to evaluate the WJ-III's construct validity (McGrew & Woodcock, 2001). A norming sample was selected that was generally representative of the U.S. population from age 24 months to age 90 years and older. Participants were selected using a stratified sampling design that controlled for gender, race, census region, and community size (McGrew & Woodcock, 2001).

The WJ-III Cognitive Battery is a nationally standardized measure of intellectual functioning. A national database provides a large-scale representative sample of the U. S. populations. In light of its large standardization sample and its reported over- sampling of African-Americans, the data from the WJ-III provide a useful database with which to employ the Spearman-Jensen hypothesis and CHC theory and to test its effects relative to reducing subgroup differences in mean IQ. Moreover, the WJ-III is the only intelligence test whose theoretical framework emanates primarily from CHC theory (Carroll, 1993; Flanagan & Ortiz, 2001; Keith, Kranzler, & Flanagan, 2001; McGrew & Woodcock, 2001).

Instrumentation

The WJ-III cognitive battery was designed to measure the intellectual abilities described in Cattell-Horn-Carroll theory of intelligence (see pages 17 through 23 of this manuscript). Figure 3-1 visually illustrates the CHC theoretical basis of the WJ-III. Stratum I includes the most specific or narrow abilities. Stratum II arises from a grouping of these narrow Stratum I cognitive abilities. These include fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and processing speed. Stratum III, the general factor, g, is derived from a combination of Strata I and II, is called General Intellectual Ability (McGrew & Woodcock, 2001). Although the WJ-III uses all three Strata as part of its underlying framework, greatest emphasis and coverage are placed on Stratum II of the CHC factors because of their reliability and direct contribution to General Intellectual Ability (McGrew & Woodcock, 2001). The aforementioned not withstanding, each Stratum I test included in the battery was a single measure of narrow abilities (McGrew & Woodcock, 2001). That is, each subtest contains substantial test specificity.

Broad factors on the WJ-III are theoretical constructs that are well-defined and based on extensive internal and external validity evidence (McGrew & Woodcock, 2001). Clusters on the WJ-III are derived from two or more subtests (McGrew & Woodcock, 2001). WJ-III clusters for both the standard and extended Cognitive Batteries include General Intellectual Ability, Verbal Ability, Thinking Ability, and Cognitive Efficiency. The first seven subtests on the standard battery contribute to the General Intellectual Ability cluster. On the Achievement Battery, the Broad Reading cluster is comprised of measures of Letter-Word Identification, Math Fluency, and Passage Comprehension. The Broad Math cluster is comprised

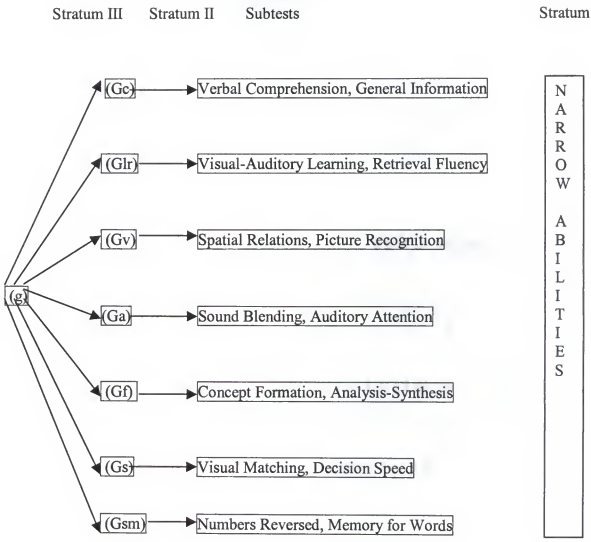

Figure 3-1. WJ-III Tests of Cognitive Abilities as it Represents CHC Theory

of measures of Calculation, Math Fluency, and Applied Problems. The broad written language cluster is comprised of measures of Spelling, Writing Fluency, and Writing Samples.

Test Reliability

One purpose of this study is to compare the mean scores between African-Americans and Caucasian-Americans. Reliability of test scores is prerequisite to this issue. Thus, reliability coefficients are relevant to this discussion.

Internal consistency reliability coefficients for the WJ-III clusters were calculated using Mosier's (1943) equation and procedures. Internal consistency reliability coefficients for the WJ-III subtests were calculated using either the split-half procedure or the Rasch analysis procedures. Split-half procedures were not appropriate for speeded tests or tests with multiple-point scored items (McGrew & Woodcock, 2001).

Median subtest internal consistency reliability coefficients for Stratum II abilities on the standard WJ-III Cognitive battery range from .81 to .94. The median reliability coefficient for the General Intellectual Ability is .97. Table 3-1 reports the median reliability coefficients for the relevant achievement tests. All median reliabilities for the achievement battery are .85 or higher. All median reliabilities for the achievement subtests examined in this study exceed .86.

Thus, the WJ-III subtests display rather high levels of internal consistency reliability. Test-retest, interrater, and alternate form reliability studies also reveal high degrees of reliability. The above reliability coefficients compare favorably with other frequently used intelligence tests.

Table 3-1

Reliability Statistics for the WJ-III Tests of Cognitive and Achievement Abilities by Combined Ages

WJ-III Factor	Battery	Median Reliability Combined Ages
General Intellectual Ability	Cognitive	.97
Stratum II		
Verbal Comprehension	Cognitive	.92
Visual-Auditory Learning	Cognitive	.86
Spatial Relations	Cognitive	.81
Sound Blending	Cognitive	.89
Concept Formation	Cognitive	.94
Visual Matching	Cognitive	.91
Numbers Reversed	Cognitive	.87

WJ-III Factor	Battery	Median Reliability Combined Ages
Broad Reading	Achievement	.94
Letter-Word Identification	Achievement	.94
Reading Fluency	Achievement	.90
Passage Comprehension	Achievement	.88
Broad Math	Achievement	.95
Calculation	Achievement	.86
Math Fluency	Achievement	.90
Applied Problems	Achievement	.93
Broad Written Language	Achievement	.94
Spelling	Achievement	.90
Writing Fluency	Achievement	.88
Writing Samples	Achievement	.87

Table 3-2

Comparison of Fit of WJ-III CHC Broad Model Factor Structure with Alternative Models in the Age 6 to Adult Norming Sample

Models	Chi-Square	df	AIC	RMSEA
WJ-III 7-Factor	13.189.16	536	13,377.16	.056 (.055-.057)
g single Factor	65,314.78	1,170	65,524.78	.086 (.085-.086)
Null Model	215,827.54	1,219	215,939.54	.153 (.153-.154)

Source: WJ-II Technical Manual.

Table 3-3

Confirmatory Factor Analysis Broad Model, g-Loadings – Age 6 to Adult Norming Sample

	Broad Factors						
Test	Gc	Glr	Gv	Ga	Gf	Gs	Gsm
Verbal Comprehension	.92						
Visual-Auditory Learning		.80					
Spatial Relations			.67				
Sound Blending				.65			
Concept Formation					.76		
Visual Matching						.71	
Numbers Reversed							.71

Source: WJ-III Technical Manual.

The test authors noted that, "The reliability characteristics of the WJ-III meet or exceed basic standards for both individual placement and programming decisions. The interpretive plan of the WJ-III emphasized the principle of cluster interpretation for most important decisions. Of the median cluster reliabilities reported, most are .90 or higher.... Of the median test reliabilities reported, most are .80 or higher and several are .90 or higher" (McGrew & Woodcock, 2001, p. 48).

Salvia and Ysseldyke (1991) recommend certain standards relative to test reliabilities coefficients in high-stakes testing. They consider reliability coefficients of .90 or higher as critical for making important educational and diagnostic decisions (e.g., special education placement). Reliability coefficients at or above .80 are thought to be important for tests used to make screening decisions. Reliability coefficients below .80 are thought to be insufficient to make decisions about an individual's test performance (McGrew & Flanagan, 1998). Reliability coefficients for WJ-III cluster scores meet these criteria.

Test Validity

As previously indicated, test validity is considered to be found in empirical evidence and theory that support the actual and potential uses of tests, including their consequences (American Educational Research Association, et al., 1999). The WJ-III Technical Manual provides information on four types of validity: (a) test content, (b) developmental patterns of scores, (c) internal structure, and (d) relationships with other external variables (McGrew & Woodcock, 2001). The WJ-III Technical Manual addresses the consequence of score interpretation and use tangentially in that these issues largely are the responsibility of test users not, test producers.

Each subtest was included in the cognitive battery because confirmatory factor analyses (tables 3-2 and 3-3) revealed almost all of them loaded exclusively on a single factor (McGrew & Woodcock, 2001). This evidence suggested limited construct-irrelevant variance on the cognitive tests (McGrew & Woodcock, 2001, p. 101).

Several studies that examine relationships between General Intellectual Ability the WJ-III and other intelligence tests (e.g., Wechsler scales, the Differential Abilities Scales, and the Standford-Binet Intelligence Scale: Fourth Edition) demonstrate correlations consistently in the .70s across samples (McGrew & Woodcock, 2001). These concurrent validity data are comparable to data reported in the most frequently used intelligence tests (e.g., Wechsler scales and the Standford-Binet Intelligence Scale: Fourth Edition). The results of these studies are reported in tables 4-5 through 4-9 of the Technical Manual (McGrew & Woodcock, 2001).

The WJ-III Technical Manual reports achievement battery data for content, development, construct, and concurrent validity. The data indicate the achievement battery measures academic skills and abilities similar to those measured by other frequently used achievement tests (e.g., Wechsler Individual Achievement Test, 1992 and the Kaufman Test of Education Achievement, 1985).

Test Fairness

According to the authors, the WJ-III was designed to attenuate test bias associated with gender, race, or Hispanic origin. Item development was conducted using recommended experts' viewpoints as to potential item bias and sensitivity. The test authors do not indicate how the experts were selected. That is, no information was provided regarding necessary criteria to be considered an expert. Items were modified or

eliminated when statistical analyses upheld an expert's assertion that an item was potentially unsuitable.

Rasch statistical methods were used to determine the fairness of WJ-III item functioning for all racial, ethnic, and gender groups. The Comprehension-Knowledge (Gc) subtests (i.e., Verbal Comprehension and General Information) were studied intensely for item fairness because the majority of items identified by experts as potentially unsuitable were from this cluster.

Factor Analysis

The authors conducted a factor-structure invariance study by male/female, white/non-white, and Hispanic/non-Hispanic groups. The resultant data suggest WJ-III scores are not biased against members of these groups. Overall, the WJ-III seems to assess the same cognitive constructs across racial, ethnic, and gender groups (McGrew & Woodcock, 2001). The test authors report the factor structure of the WJ-III to be the same for relevant subgroups (tables 3-2 and 3-3). They conducted factor invariant analysis the following procedures:

> Using Horn, McArdle, and Mason's (1983) suggestion that 'configural invariance' – tests loading on the same factors across groups – is the most realistic and recommended test of factor structure invariance, group CFA was completed for White/non-White group drawn from the standardization sample (age 6 and older). The same factor model was specified for both sub-groups (e.g., White and non-White), with the same factors and the same pattern of factor loadings. Such an analysis tests for configural invariance across groups. Using the RMSEA fit statistic (with a 90% confidence interval) to evaluate the analysis, the WJ-III broad factor model was found to be a good fit in the White/non-White (RMSEA = .039; .038 to .039) analysis. (McGrew & Woodcock, 2001, p. 100)

Carroll (1993) found that the CHC theoretical model is uniform across race. Overall, the WJ-III authors' confirmatory factor analytic studies suggest the WJ-III is largely invariant across race and reflects a "fair" formulation for both groups. However, additional

invariance analyses will be conducted to determine whether loadings for each test factor differ between African-Americans and Caucasian-Americans.

Procedures

Consent to conduct the study was obtained from the University of Florida's Institutional Review Board. Dr. Thomas Oakland obtained the WJ-III standardized data from Drs. Richard Woodcock and Kevin McGrew. Dr. Woodcock was asked to supply the following information from the WJ-III: standard scores on the cognitive battery from the standardization sample by ethnicity, gender, SES, and mean IQs of all participants. The letter requesting use of the standardization sample data served as the informed consent document.

No potential risks accrue to study participants because the data are archival and do not contain any personally identifying information. Demographic information on race, gender, and SES was acquired from the data set.

Methodology

The most widely used method to measure agreement between factor structures across groups is the congruence coefficient, r_c (Kamphaus, 2001). The congruence coefficient is an index of factor similarity and is interpreted similar to a Pearson correlation coefficient (Jensen, 1998). "A value of r_c of +.90 is considered a high degree of factor similarity; a value greater than + .95 is generally interpreted as practical identity of the factors. The r_c is preferred to the Pearson r for comparing factors, because the r_c estimates the correlation between the factors themselves, whereas the Pearson r gives only the correlation between the two column vectors of factor loadings" (Jensen, 1998, p. 99). The congruence coefficient was used to measure agreement between the factor structures for African-Americans and Caucasian-Americans.

Multivariate analysis of variance (MANOVA) was used to test hypotheses regarding whether mean scores differ based on race. Principal component factor analysis and the congruence coefficient test were used to determine whether the factor structures of the two groups differ.

Mean differences among racial/ethnic groups obtained from different studies or different intelligence tests are averaged best when mean differences are stated in units of the averaged standard deviation within the racial/ethnic groups. The sigma difference or effect size (d) test allows direct comparisons of mean differences irrespective of the scale of measurement or the quality measured (Jensen, 1998). The procedure is similar to Cohen's d (Cohen, 1988) and the use of z score analyses. The sigma difference determines the significance of the results. Thus, the sigma difference or effect size (d) test was used to determine whether the expected reduced mean score difference between African-Americans and Caucasian-Americans differs significantly from 15 points. This statistic permits direct comparisons of mean difference regardless of the original scale of measurement (Jensen, 1998). That is, the mean difference observed on the WJ-III can be compared directly to the traditionally observed mean difference of 15 points. The sigma difference or effect size metric also was used to determine whether smaller mean differences would be evident on Stratum II compared to Stratum III factors.

An understanding of the practical importance of significant differences requires information regarding effect sizes. The Omega Hat Squared statistic should be used with sample sizes larger than one thousand. Cohen (1988) suggests small effect sizes occur between .01 and .05, moderate effect sizes occur between .06 and .14, and large effect sizes occur at or above .15.

Pearson correlations between general intelligence and the nine academic achievement subtests and three broad clusters (Table 4-1 shows the subtests and clusters) were obtained for African-Americans and Caucasian-Americans. The achievement subtests are those that contribute to the three clusters of Broad Reading, Broad Math, and Broad Written Language. Correlation coefficients were examined for significance using Pearson's correlation coefficient test. The Fisher Z transformation (not to be confused with z score analysis) was used to determine whether the correlation coefficients between the two groups differed.

The independent variables in this study are racial/ethnic group: African-Americans and Caucasian-Americans. The dependent variables are IQs and standard scores for each group on Strata I, II, and III for both the standard and achievement batteries.

CHAPTER 4
RESULTS

Principal Component Factor Analysis

Principal component factor analysis was conducted on the Strata II and III factors for African-Americans and Caucasian-Americans. Principal component g loadings were obtained (Table 4-8). Correlation of congruence (r_c) was conducted to determine whether g loadings were similar between African-Americans and Caucasian-Americans. The results of the analyses reveal a congruence coefficient, r_c of .99. It indicates the factor structure of the WJ-III does not differ for African-Americans and Caucasian-Americans. In fact, the factor structures are almost identical for the two groups.

MANOVA

A MANOVA was computed using race (African-Americans and Caucasian Americans) as the nominal, independent, or factor variables. IQs on the WJ-III Stratum II and Stratum III were used as the dependent variables (Tables 4-1 through 4-6). The MANOVA tested whether mean scores on the WJ-III Strata II and III, are higher for Caucasian-Americans than African-Americans. Caucasian-American obtained higher IQs than African-Americans ($F = 44.8$; $P < .001$). Strata II and III scores are significantly higher for Caucasian-Americans than for African-Americans. The magnitude of the mean difference is 11.3 on the General Intellectual Ability factor, 13.4 on the Verbal Comprehension, 5.2 on the Visual-Auditory Learning, 5.0 on the Spatial Relations, 9.9 on the Sound Blending, 9.8 on the Concept Formation,

2.9 on the Visual Matching, and 6.2 on the Numbers Reversed tests. Univariate findings indicate all mean differences are significant at the $P < .001$ or better (Tables 4-2 through 4-6).

Effect Size Test for Large Samples

Cohen (1988) suggests small effect sizes occur between .01 and .05, moderate effect sizes occur between .06 and .14, and large effect sizes occur at or above .15. The Omega Hat Squared effect size (used with large sample sizes) to determine the importance of the differences observed between the two groups is .08 for General Intellectual Ability, a figure considered to be a moderate effect size based on Cohen's (1988) criteria. Additionally, moderate effect size differences of .12 for Verbal Comprehension, .07 for Sound Blending, and .06 for Concept Formation were evident. Small effect sizes of .02 for Visual-Auditory Learning, .02 for Spatial Relations, .02 for Numbers Reversed, and .01 for Visual Matching were obtained. Strong effect sizes are considered of practical significance and weak effect sizes suggest limited practical significance.

Sigma Difference Test

The sigma difference test was used to determine whether the mean score difference on the WJ-III General Intellectual Ability factor between Caucasian-Americans and African-Americans is less than 15 points. The mean General Intellectual Ability score difference between the two groups of 11.3 points results in a sigma difference of .81 (Table 4-7). Meta-analytic studies reveal an observed overall mean sigma difference is 1.08, with a standard deviation of 0.36 (Jensen, 1998). Given a normal distribution, about two-thirds of the mean differences between Caucasian-Americans and African-Americans are between 0.72 and 1.44. Considering a 15-point standard deviation,

approximately two-thirds of the mean differences between the two groups are between ten and twenty IQ points. A sigma difference of .81 is substantially below the overall typical mean sigma difference of 1.08 and reflects a reduction of 25 %. Nonetheless, a sigma difference of .81 is within the range of what was obtained in the meta-analysis.

Subtracting 1.08 from .81 results in an effect size change of -.27, a figure considered to be an extremely large effect size using Cohen's (1988) criteria. Overall, the results reveal mean IQ differences between Caucasian-Americans and African-Americans are significantly smaller on the WJ-III than 15 points. Once again, the sigma difference or effect size (d) test allows direct comparisons of mean differences irrespective of the scale of measurement or the quality measured (Jensen, 1998).

The sigma difference test was used to determine whether mean differences between African-Americans and Caucasian-Americans will be smaller on Stratum II than on Stratum III. Compared to the degree of difference between African-Americans and Caucasian-Americans on the General Intellectual Ability factor, mean differences are smaller on all Stratum II factors but one (Verbal Comprehension) (Table 4-6). Mean differences between Verbal Comprehension, Visual-Auditory Learning, Sound Blending, Concept Formation, Spatial Relations, Visual Matching, and Numbers Reversed and Stratum III: General Intellectual Ability are significant at $p < .001$. Additionally, moderate Omega Hat Squared effect sizes of .12 for Verbal Comprehension, .07 for Sound Blending, and .06 for Concept Formation were evident. Small effect sizes of .02 for Visual-Auditory Learning, .02 for Spatial Relations, .01 for Visual Matching, and .02 for Numbers Reversed were noted.

A mean difference of 13.4 on the Verbal Comprehension subtest is significant at the $p < .001$ (with an Omega Hat Squared effect size of .12). This difference is both

larger than the 11.3 difference observed on General Intellectual Ability and is in the opposite direction of the stated hypothesis. Its effect size .12, is considered to be moderate.

Sigma difference changes (Table 4-7) among the seven broad factors and General Intellectual Ability reveal large effect sizes on Verbal Comprehension (.98 -.81 = .17, but in the opposite direction as that hypothesized), Visual-Auditory Learning (.38 - .81 = -.43), Spatial Relations (.36 - .81 = -.45), Visual Matching (.20 - .81 = -.61), and Numbers Reversed (.40 - .81 = -.41). Moderate effect size changes are found on Sound Blending (.70 - .81 = -.11) and Concept Formation (.68 - .81 = -.13). Thus, compared to racial differences on General Intellectual Ability, differences between African-Americans and Caucasian-Americans are less on the following subtests: Visual-Auditory Learning, Spatial Relations, Visual Matching, and Numbers Reversed. The magnitude of racial differences on General Intellectual Ability does not appreciably differ from those on Sound Blending and Concept Formation. Differences between African-Americans and Caucasian-Americans are moderately larger on Verbal Comprehension than on the general intelligence.

Correlations Between General Intelligence and Achievement

Means (Table 4-9) and correlation coefficients r (Table 4-10) were obtained for General Intellectual Ability and each academic achievement subtest that comprise the Broad Reading, Broad Math, and Broad Written Language factors. Pearson correlations indicate all of the subtests correlate significantly with General Intellectual Ability for both groups, $P < .001$ (Table 4-10).

Fisher's Z transformation was used to compare correlations between General Intellectual Ability and Broad Reading, Broad Math, and Broad Written Language as well

as for each academic achievement subtest that comprise these three Broad factors for African-Americans and Caucasian-Americans. Applying Fisher's statistic, all z scores are less than .001 and are not significant at alpha = .05. Thus, correlations between general intelligence and the 12 academic achievement scores do not differ significantly for African-Americans and Caucasian-Americans.

Table 4-1

WJ-III Cognitive and Achievement Batteries Codes

GIA – General Intellectual Ability
Gc – Verbal Comprehension
Glr – Visual-Auditory Learning
Gv – Spatial Relations
Ga – Sound Blending
Gf – Concept Formation
Gs – Visual Matching
Gsm – Numbers Reversed
Reading – Broad Reading
Letter-Word Identification
Reading Fluency
Passage Comprehension
Math – Broad Math
Calculation
Math Fluency
Applied Problems
Written Language – Broad Written Language
Spelling
Writing Fluency
Writing Samples

Table 4-2

Box's Test of Equality of Covariance Matrices – Homogeneity of the Variance

Box's M	153.7
F	4.2
df1	36
df2	1415586
Sig.	.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.
Design: Intercept + Race

Table 4-3

Bartlett's Test of Sphericity

Likelihood Ratio	.000
Approx. Chi-Square	9288.8
df	35
Sig.	.000

Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.
Design: Intercept + Race

Table 4-4

Multivariate Tests of Significance Effect for Group

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	Pillai's Trace	.99	25351.4	8.0	2146.0	.000	.99	202811.4	1.00
	Wilks' Lambda	.01	25351.4	8.0	2146.0	.000	.99	202811.4	1.00
	Hotelling's Trace	94.5	25351.4	8.0	2146.0	.000	.99	202811.4	1.00
Race	Pillai's Trace	.14	44.8	8.0	2146.0	.000	.14	358.4	1.00
	Wilks' Lambda	.86	44.8	8.0	2146.0	.000	.14	358.4	1.00
	Hotelling's Trace	.17	44.8	8.0	2146.0	.000	.14	358.4	1.00

Computed using alpha = .01
Exact statistic
The statistic is an upper bound on F that yields a lower bound on the significance level.
Design: Intercept + Race

Table 4-5

Levene's Test of Equality of Error Variances

	F	df1	df2	Sig.
General Intellectual Ability	6.9	1	2153	.009
Verbal Comprehension	9.1	1	2153	.003
Visual-Auditory Learning	1.2	1	2153	.265
Spatial Relations	2.1	1	2153	.148
Sound Blending	20.4	1	2153	.000
Concept Formation	.14	1	2153	.709
Visual Matching	2.7	1	2153	.101
Numbers Reversed	.93	1	2153	.335

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
Design: Intercept + Race

Table 4-6

Univariate Tests

Dependent Variable	Mean IQs - CA	Mean IQs -AA	Mean Difference	F	Sig.	Power	$W^{\wedge 2}$
General Intellectual Ability	104.3	93.0	11.3	196.2	.000	1.00	.08
Verbal Comprehension	104.2	90.8	13.4	284.8	.000	1.00	.12
Visual-Auditory Learning	103.7	98.5	5.2	42.3	.000	1.00	.02
Spatial Relations	102.6	97.5	5.1	39.1	.000	1.00	.02
Sound Blending	104.2	94.3	9.9	148.7	.000	1.00	.07
Concept Formation	103.5	93.7	9.8	140.4	.000	1.00	.06
Visual Matching	101.5	98.6	2.9	12.6	.000	.994	.01
Numbers Reversed	103.0	96.8	6.2	51.1	.000	1.00	.02

The F tests the effect of Race: CA = Caucasian-Americans; AA = African-Americans. The test is based on the linearly independent pairwise comparisons among the estimated marginal means.

* The mean difference is significant at alpha = .05.

$W^{\wedge 2}$ = Omega Hat Squared

Table 4-7

Sigma Difference - Direct Comparison of Changes in Effect Size for the GIA and Each Stratum II Subtest

Dependent Variable	Sigma Difference Effect Size	Change in Effect Size	Cohen's "Error Effect" Guidelines
General Intellectual Ability	.81	---	---
Verbal Comprehension	.98	.166	Large Effect
Visual-Auditory Learning	.36	-.434	Large Effect
Spatial Relations	.36	-.448	Large Effect
Sound Blending	.70	-.105	Moderate Effect
Concept Formation	.68	-.125	Moderate Effect
Visual Matching	.21	-.605	Large Effect
Numbers Reversed	.40	-.410	Large Effect

Sigma difference standardized scale permits direct comparison of mean differences regardless of the original scale of measurement

Table 4-8

Principal Component Matrix

WJ-III Test	CA: g loading	AA: g loading	ESS: g loading	Mean IQ Difference
General Intellectual Ability	.97	.96	.97	11.307
Verbal Comprehension	.78	.79	.79	13.423
Visual-Auditory Learning	.76	.74	.76	5.201
Spatial Relations	.53	.50	.55	5.036
Sound Blending	.63	.62	.65	9.938
Concept Formation	.79	.79	.80	9.828
Visual Matching	.59	.60	.59	2.928
Numbers Reversed	.69	.61	.68	6.236

First Principal Component g-factor loadings calculated for the GIA and the seven subtests
CA = Caucasian-Americans
AA = African-Americans
ESS = Entire standardization Sample

Table 4-9

Descriptive Statistics – Caucasian-Americans and African-Americans

	Caucasian-Americans			African-Americans		
	Mean	SD	N	Mean	SD	N
General Intellectual Ability	104	14.3	1978	93.3	13.0	401
Broad Reading	103	14.5	3142	96.3	13.7	523
Broad Math	102	14.6	3683	96.5	14.1	585
Broad Written Language	105	14.4	3359	97.7	14.6	543
Letter-Word Identification	103	15.1	3913	96.6	15.9	617
Reading Fluency	103	14.6	3174	96.0	13.3	533
Passage Comprehension	102	15.4	3910	97.1	15.1	616
Calculation	101	15.8	3826	97.6	15.4	615
Math Fluency	100	14.4	3723	98.0	15.4	596
Applied Problems	102	15.4	3856	94.9	13.7	615
Spelling	103	14.6	3851	98.8	15.6	617
Writing Fluency	103	15.9	3370	95.1	15.6	543
Writing Samples	103	15.5	3802	98.4	16.0	612

Table 4-10

Pearson Correlations between General Intelligence and Academic Achievement for African-Americans and Caucasian-Americans

	Caucasian-Americans				African-Americans			
	N	General Intellectual Ability	Sig.	Sum of Squares	N	General Intellectual Ability	Sig.	Sum of Squares
Broad Reading	1851	.720	.001	256147	365	.719	.001	46678
Broad Math	1899	.642	.001	235065	381	.642	.001	42692
Broad Written Language	1788	.645	.001	229831	357	.653	.001	42601
Letter-Word Identification	1975	.617	.001	257022	401	.647	.001	52649
Reading Fluency	1851	.604	.001	220398	365	.593	.001	37557
Passage Comprehension	1975	.574	.001	236425	400	.591	.001	42377
Calculation	1960	.487	.001	193476	397	.477	.001	35802
Math Fluency	1901	.460	.001	171956	381	.491	.001	38706
Applied Problems	1974	.578	.001	236943	399	.643	.001	42727
Spelling	1972	.552	.001	221926	399	.542	.001	42607
Writing Fluency	1789	.499	.001	187895	357	.556	.001	37694
Writing Samples	1953	.508	.001	208124	396	.527	.001	42952

Table 4-11

Fisher Z Transformation: z-test for Independent Correlations between Caucasian-Americans and African-Americans for General Intelligence and Academic Achievement

	Caucasian-Americans		African-Americans		Both Groups
	Pearson's r General Intellectual Ability	Fisher Z Transformation	Pearson's r General Intellectual Ability	Fisher Z Transformation	z-scores
Broad Reading	.72	.91	.72	.91	0000
Broad Math	.64	.76	.64	.76	0000
Broad Written Language	.65	.77	.65	.78	.0004
Letter-Word Identification	.62	.72	.65	.77	-.0010
Reading Fluency	.60	.70	.59	.69	.0003
Passage Comprehension	.57	.66	.59	.68	-.0005
Calculation	.49	.53	.48	.52	.0003
Math Fluency	.46	.50	.49	.54	-.0008
Applied Problems	.58	.66	.64	.77	-.0022
Spelling	.55	.62	.54	.60	.0003
Writing Fluency	.50	.55	.56	.63	-.0017
Writing Samples	.51	.56	.53	.58	-.0004

CHAPTER 5
DISCUSSION

Two primary imperatives motivated this research: one theoretical and one practical. The first imperative provided the theoretical underpinnings for the study and involved testing the Spearman-Jensen hypothesis in light of the recently developed and comprehensive set of data from the WJ-III, a test developed to be consistent with CHC theory. The second imperative was to provide data on the mean score differences between Caucasian-Americans and African-Americans on the recently published WJ-III measure of cognitive ability and academic achievement.

Prior to testing the Spearman-Jensen hypothesis, data revealed the factor structure of the WJ-III to be consistent for African-Americans and Caucasian-Americans. This finding allows one to test the Spearman-Jensen hypothesis with greater confidence that the data reflect a similar construct of intelligence. In view of the Spearman-Jensen hypothesis, African-Americans were expected to obtain lower IQs than Caucasian-Americans. The results of this research indicate African-Americans continue to evidence lower mean IQs than Caucasian-Americans. As hypothesized, African-Americans scored lower on the General Intellectual Ability factor and on all broad factors. Additionally, on this intelligence test comprised of both broad and specific factors associated with the hierarchical approach of CHC Theory, a significantly smaller mean racial difference was

displayed (i.e., 11 points on the WJ-III) when compared to the traditionally observed 15 points.

In practice, a difference of four IQ points can influence whether a child is considered gifted, mentally handicapped, and learning disabled. A difference of four IQ points also may impact the disproportionate representation of African-Americans in other specialized programs. On intelligence tests where African-Americans average scores are four points less than on the WJ-III, there is a greater likelihood they will be over-represented in mentally handicapped and developmentally delayed programs and underrepresented in gifted programs.

Smaller Differences on Broad Factors than on g

In light of the fact broad factors have smaller g loadings than the General Intellectual Ability factor, mean differences between African-Americans and Caucasian-Americans were expected to be smaller on these broad factors than on the General Intellectual Ability factor. This hypothesis was supported. Mean IQ differences were smaller on six of the seven broad factors. Sigma difference changes between the seven broad factors and General Intellectual Ability reveal large effect sizes for Visual-Auditory Learning, Spatial Relations, Visual Matching, and Numbers Reversed. Moderate effect sizes were evidence for Sound Blending and Concept Formation (Table 4-10). Thus, as hypothesized, differences between African-Americans and Caucasian-Americans generally are less on the seven broad factors than on General Intellectual Ability. The Verbal Comprehension factor does not display this trend. Mean score differences are larger on Verbal Comprehension than on the General Intellectual Ability factor.

The Spearman-Jensen hypothesis suggests mean IQ differences between African-Americans and Caucasian-Americans occur as a function of the tests' g loadings. As

previously discussed, tests of broad and narrow ability are comprised of g as well as factors specific to each test. Specificity refers to the proportion of a test's true score variance that is unaccounted for by a common factor such as g (Jensen, 1998). On most WJ-III Cognitive Battery subtests, more than 50% of the variance of each subtest is specific to that subtest (Table 4-8). As such, its sources of variance are partly comprised of g and partly comprised of qualities other than g (Jensen, 1998).

IQ differences between African-Americans and Caucasian-Americans should be smaller on tests with larger specificity because of their lower g loadings. That is, the larger a test's specificity, the smaller the mean IQ difference one should find between African-Americans and Caucasian-Americans. Overall, the results support the Spearman-Jensen hypothesis. One possible reason for the Verbal Comprehension exception is that in addition to the high g loading found on the Verbal Comprehension subtest, the test possesses rather high cultural loadings (Flanagan & Ortiz, 1998). The test authors' noted that most of the test items that raised concerns regarding bias were from the comprehension-knowledge tests (McGrew & Woodcock, 2001). Therefore, it appears further investigations regarding the fairness of this subtest should contemplated.

Similar Factor Structures for Both Groups

The findings of this study support the test authors' assertion that the factor structures of the WJ-III for Caucasian-Americans and African-Americans are consistent. Confirmatory factor analysis reveals a comparable factor model, with the same factors, and nearly identical directional pattern of factor loadings for both groups on the cognitive battery (McGrew & Woodcock, 2001). Moreover, findings show consistent g-loading scores for both groups on the eight cognitive battery variables.

The congruence coefficient, r_c for African-Americans and Caucasian-Americans on Strata II and III of the WJ-III is .99. Thus, the factor structures of Strata II and III are essentially identical for both groups. Clearly, g accounts for similar amounts of variance in IQ for Caucasian-Americans and African-Americans on the WJ-III. These results support the test authors' findings that the WJ-III measures the same factors for Caucasian-Americans and African-Americans. The study also supports Carroll's (1993) finding that CHC is essentially invariant across racial/ethnic groups.

Correlations between general intelligence and Broad Reading, Broad Math, and Broad Written Language and the subtests that comprise these factors are similar for Caucasian-Americans and African-Americans. All correlations are statistically significant at the $p < .01$, thus adding to evidence that the WJ-III is measuring the same construct for both groups. These findings also support the test authors' contention that the WJ-III measures the same factors for African-Americans and Caucasian-Americans.

Significance of g

The findings of this study support the Spearman-Jensen hypothesis and Spearman's two-factor theory of intelligence to a greater degree than CHC theory. Support for Spearman's two-factor theory is somewhat surprising because CHC theory considers intelligence to be hierarchical rather than bi-factorial. A major component of the theory is that several broad and specific factors, measurably different from g, are instrumental in determining intelligence test scores. According to proponents of CHC theory, broad and specific factors are linearly independent. However, on the WJ-III cognitive battery, subtests contain substantial g loadings. The g loadings for standard battery broad factors are greater than .55 and average .72. G loadings for the different Stratum II factors on the WJ-III are sufficiently high to suggest they primarily measure

the principal component, g. Therefore, the subtests may not be entirely linearly independent. Thus, the WJ-III is viewed as a highly g-loaded measure.

In light of the Spearman-Jensen hypothesis, one expects to find substantial mean IQ differences between African-Americans and Caucasian-Americans on highly g-loaded tests. The results of this research are consistent with this and a two-factor understanding of intelligence, but not entirely consistent with a hierarchical understanding of intelligence.

Despite the hierarchical nature of CHC theory, broad factors, although considered different from g in the theory, substantially add to the variance associated with intelligence test performance and thus may be more similar than dissimilar from g. Thus, Stratum II broad factors appear closely related to and highly correlated with a general factor. For example, although fluid intelligence is considered a broad factor under CHC theory, it is almost indistinguishable from g (Gustafsson, 2001).

As previously noted, the Spearman-Jensen hypothesis suggests mean subgroup IQ differences are a function of variance associated with g and little else. The finding of substantial mean IQ differences between African-Americans and Caucasian-Americans on the WJ-III cognitive battery general intellectual and seven subtest factors suggests the instrument largely measures g. That is, scores on the WJ-III cognitive battery subtests are highly influenced by a general factor of ability. Recall g loadings for the standard battery broad factors average .72. Perhaps the WJ-III achievement battery, as a Stratum I factor, better represents specific and narrow abilities. That is, the cognitive battery by itself does not entirely reflect CHC theory of specific and narrow factors as important in intelligence. Rather, it is the combination of the cognitive and achievement batteries that best reflects

CHC theory. As a consequence, the measurement of the cognitive abilities requires the use of the two tests that comprise the entire battery.

Consequential Validity Perspective

To reiterate, this study was not conducted to test the reliability or validity of the WJ-III. The test authors conducted substantial analyses of the reliability and validity of the instrument. Moreover, they provide ample evidence that supports the utility of the test in school settings (McGrew & Woodcock, 2001). This study also does not indicate the instrument is biased against African-Americans or any group. In fact, in view of the 11-point mean difference between Caucasian-Americans and African-Americans on the WJ-III, this may be the intellectual measure of choice for use with African-Americans.

A more global area of concern addressed by this study is whether there are reductions in mean IQ differences between African-Americans and Caucasian-Americans in light of the Spearman-Jensen hypothesis and CHC theory. Clearly, a reduction of 4 mean IQ points is important to the educational programming of African-American students. A question raised by this study is whether the testing process is as fair possible for minorities when test users are not provided information regarding mean IQ differences for relevant subgroups. The answer appears patently obvious. Knowledge of mean IQ differences can substantially impact the testing process and educational placement of minority students. The testing process becomes less than fair when test users are unaware of mean IQ differences and cannot use this knowledge to apply good judgment in the proper selection and administration of tests.

Much of the underlying framework for this section was based on information provided by The Standards (American Educational Research Association, et al., 1999) regarding test scores and test score use as a function of validity. According to The

Standards, "evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use" (American Educational Research Association, et al., 1999, p. 83).

> When tests are used as part of decision-making that has high-stakes consequences for students, evidence of mean score differences between relevant subgroups should be examined, where feasible. When mean differences are found between subgroups, investigations should be undertaken to determine that such differences are not attributable to construct underrepresentation or construct irrelevant error. *Evidence about differences in mean scores and the significance of the validity errors should also be considered when deciding which test to use.* (U.S. Department of Education, Office for Civil Rights, 2000, p. 45; emphasis added)

Based on the above statements, the position herein is that when two distinct intelligence tests are similarly reliable and possess comparable statistical qualities, the more socially valid test is the measure with the smaller mean IQ difference between relevant subgroups groups. These groups may differ by race, ethnicity, first language, or gender. Using tests with smaller mean IQ differences between relevant subgroups groups is particularly germane when the measures are used with the lower scoring group.

Test Selection and Administration

Practitioners frequently individually determine which intelligence test they administer. Thus, to a degree, practitioners' philosophical orientations can determine students' potential to score lower or higher on intelligence tests. Judgments regarding test selection and administration when mean IQ differences occur between two statistically sound instruments will influence educational decision making. Use of an intelligence test that more favorable reflects the scores of traditionally lower performing subgroups can decrease the consequential impact and increase the social validity of test scores. For example, an African-American child who obtains an IQ of 69 on the WISC-III may achieve an IQ of 73 on the WJ-III. An IQ of 69 on the WISC-III has greater

potential to lead to placement in a program for mentally handicapped students than the WJ-III score of 73. IQs remain of valuable in education and society. An IQ of 130 may lead to placement in a gifted program, whereas an IQ of 126 likely will not. The consequences of differences in IQ among racial/ethnic subgroups are of substantial importance. These mean differences likely reduce problems associated with the disproportionate representation of some minorities in gifted and special education programs. Test developers are encouraged to publish data relative to mean subgroup differences.

Bearing in mind the significance of the consequential perspective of test validity, there are considerable consequences related to the testing African-Americans. As a result, decisions should be made with respect to whether administering intelligence tests to African-American students offer sufficient positive outcomes to outweigh the negative outcomes associated with test use.

To illustrate, for approximately 10 years psychologists in the state of California were not allowed to use intelligence tests when evaluating students for mentally handicapped programs. During the prohibition, a modest increase was found in the proportion of African-American students in California placed in special education programs. The proportions placed in mentally handicapped and developmentally delayed programs decreased, but the proportion placed in programs for students with learning disabilities increased (Morison, White, & Feuer, 1996).

Some wonder why we should be concerned about disproportionate representation in special education programs when these programs provide students' additional assistance and rights to an individualized education program (Donovan & Cross, 2002). A student must be labeled with a disability, indicative of some type of deficiency to meet

criteria for special education. Although the label may lead to extra assistance, it also often brings reduced expectations from the teacher, child, and perhaps parents. Of course, children who experience significant difficulty learning without special education support should receive such support. However, both the need for, and benefit of, such assistance should be determined before the label is imposed (Donovan & Cross, 2002).

Since the ratification of the Public Law 94-142 requiring states to educate all students with disabilities, children from some racial/ethnic groups receive special education services in disproportionate numbers (Donovan & Cross, 2002). The pattern of disproportionate representation is not evident in low-incidence handicaps (e.g., deaf, blind, orthopedic impairment, etc.) diagnosed by medical professionals and observable external to the school context (Donovan & Cross, 2002). As previously noted, disproportionate representation is most pronounced in the mentally handicapped and developmentally delayed classifications. Minorities are also underrepresented in gifted programs. Again, as formerly noted, placement in special education often occurs subsequent to some type of intelligence testing.

Mentally handicapped and developmentally delayed classifications are considered to carry pejorative labels in most social and educational circles. Therefore, the question is raised regarding whether, in instances of mental handicap and developmentally delayed labeling, the disadvantages associated with intelligence testing outweigh the advantages. The California data suggest African-American children who experience educational deficits will receive special education services in less pejorative programs and without the use of intelligence tests. Members of minority groups who argue against the use of intelligence tests likely will be supportive of testing and special education processes that are effective and serve to support minority children without using unflattering labels.

The Importance of Intelligence Tests

Advantages associated with the use of intelligence testing on occasion may outweigh the disadvantages. Intelligence tests, as they are currently designed, significantly impact society. In American society, good social judgment, reasoning, and comprehension are highly regarded. Society values all of the important measurable characteristics that correlate with IQ. Intelligence is correlated with income, SES, educational attainment, social success, and political power (Sattler, 1988). Additionally, intelligence tests provide information about a student's strengths and weaknesses. Intelligence testing is a highly efficient and economical means of predicting scholastic achievement and academic potential. IQs help measure a student's ability to compete academically and socially. Thus, intelligence is extremely important because IQ more than any other comparable score reveals differences in the noted important areas (Jensen, 1998). Therefore, although ending intelligence testing is unwarranted, perhaps the use of supplemental measures more relevant to the ecological environment of students will be beneficial.

Supplementing or Supplanting Intelligence Tests?

Intelligence tests measure verbal, abstract, and concept formation abilities, and predict success in school, all of which are important in industrialized societies. However, intelligence tests are not the only important measure of characteristics a society needs in its people to survive. Qualities such as motivation, persistence, concentration, and interpersonal skills are all important to successful living. Intelligence tests are pervasively used in psychoeducational assessment (Ortiz, 2000) and considerably impact students' diagnoses, interventions, and special educational and gifted placement. One can understand why individuals and minority groups who are disproportionately represented

in some programs and who do not qualify for many of the beneficial resources associated with high IQs are concerned about the frequent use of intelligence tests in schools.

Hilliard (1992) contends that the primary problem with intelligence testing is that the tests show an absence of instructional validity. Instructional validity refers to the nature of, or to the existence of, links between testing, assessment, placement, treatment, and instructional outcomes. That is, how do these tests benefit the student in light of research showing tracking and special education placement are of little help in remediating academic problems (Taylor, 1989).

Users of intelligence tests may assume that students' capacities are fixed and that attempts to compare and rank students when deciding which type of custodial care in education that they should receive is important (Hilliard, 1992). Hilliard (1992) maintains that students' cognition can be improved and that the important information to gain from evaluations are diagnostic descriptions of impediments to full functioning, not a rank order of the students. When this type of model is utilized in student evaluations; that is, the conditions that prevent full functioning, educators are better able to link test results to valid remedial instruction. This model leads to the evaluator troubling shooting the system. Evaluators must make certain their actions benefit the children for whom they are supposed to evaluate and with whom they are supposed to intervene (Hilliard, 1992).

Rather than using intelligence tests, perhaps performance and/or informal assessment measures (e.g., curriculum based and portfolio assessments) can be used to determine eligibility for some programs. While performance measures may more favorable reflect functioning of subgroups that traditionally score low on intelligence tests (Reschly & Ysseldyke, 1995), performance measures may unfavorable reflect functioning

of students who are considered gifted (Benbow & Stanley, 1996). However, use of performance measures may improve results for all students when performance competencies emphasize improvements across all achievement ranges (Braden, 1999; Meyer, 1997).

Equalizing Outcomes or Equalizing Opportunities

Braden (1999) implied that researchers and scholars should not expect to equalize educational and intelligence score outcomes for racial/ethnic groups and instead should focus their work on equalizing educational opportunities for all groups. However, economically disadvantaged populations are at greater risk for many of the causes of handicapping conditions. The etiologies associated most frequently with handicapping conditions overlap conditions associated with poverty. Economically disadvantaged populations often are more predisposed to disorders related to environmental, nutritional, and traumatic factors (U. S. Department of health and Human Services, in Westby, 1990). These factors tend to lower intelligence. As the Committee on Minority Representation in Special Education notes:

> Poverty is associated with higher rates of exposure to harmful toxins, including lead, alcohol, and tobacco, in early stages of development. Poor children are also more likely to be born with low birth weight, to have poorer nutrition, and to have home and child care environments that are less supportive of early cognitive and emotional development than their majority counterparts. When poverty is deep and persistent, the number of risk factors rises, seriously jeopardizing development. . . . In all income groups, black children are more likely to be born with low birth weight and are more likely to be exposed to harmful levels of lead. . . . While the separate effect of each of these factors on school achievement and performance is difficult to determine, substantial differences by race/ethnicity on a variety of dimensions of school preparedness are documented at kindergarten entry. (Donovan & Cross, 2002, p. ES-iii)

The above suggests researchers, scholars, and stakeholders in the use of intelligence tests with minority students should strive to do more than equalize educational opportunities.

In addition to equalizing educational opportunities, the belief herein is that equivalent efforts should be made to equalize environmental and nutritional factors that impact racial/ethnic minorities and their intelligence. Moreover, serious attempts should be made to prevent the effects of traumatic factors that may depress intellectual functioning. The aforementioned may help not only to equalize educational opportunities, but equalize intelligence and educational outcomes for the relevant minority subgroups.

Professionals who are responsible for the assessment of children who differ culturally, linguistically, or racially must realize that they are dealing with potential and very real conflicts in values. These are conflicts all who assess minority children incur, with test cultural loading, social issues, and social and consequential validity weighed on one hand, and statistical, psychological, and educational theories, practices and decisions weighed on the other. It is at this point that each individual psychologist makes philosophical decisions about whether a particular test or for that matter testing itself is appropriate (Messick & Anderson, 1970). The deciding factor always should be whether the positive consequences associated with testing will outweigh the negative consequences.

When deciding on testing and which test to administer, both statistical bias and indices of consequential bias should be considered. Recall statistical bias in testing essentially concerns the presence of construct irrelevant components and construct under-representation in tests that produce systematically lower or higher scores for subgroups of test takers. The current contention is that tests also should be considered biased when the negative consequences associated with their use outweigh the positive consequences. Consequential bias refers to the use of test scores that result in substantial disadvantages accruing to subgroups as a function of the test's predictive imprecision (e.g., on criteria

measures such as academic achievement, grades, attainment of high school diplomas and college degrees, etc). Thus, bias in this context refers to the social and educational disadvantages resulting from the use of intelligence tests. All else being equal, the intelligence test with the greater consequential bias is the test with a greater disparate mean between relevant subgroups. If, because of political, administrative, or societal reasons one must administer intelligence and other standardized tests, one must be certain to make decisions based not only test reliability and validity, but on the social consequences of test results as a function of test fairness.

In light of the findings, this study may serve as the catalyst to encourage all intelligence test publishers to supply test users with data, concerning not only factor structure differences, but data regarding mean IQ differences between various racial/ethnic groups. Political correctness should not subjugate scholarly precision.

REFERENCES

Aaron, P.G. (1997). The impending demise of the discrepancy formula. Review of Educational Research, 67, 461-502.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: Author.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. Journal of Early Adolescence,14, 3, 328-344.

Artiles, A.J., & Trent, S.C. (1994). Overrepresentation of minority students in special education: A continuing debate. Journal of Special Education, 27, 410-437.

Benbow, C.P., & Stanley, J.C. (1996). Inequity in equity: How "equity" can lead to inequity for high-potential students. Psychology, Public Policy, and Law, 2, 249-292.

Bracken, B.A. (1985). A critical review of the Kaufman Assessment Battery for children (K-ABC). School Psychology Review, 14, 21-36.

Bracken, B.A., & McCallum, R.S. (1998). Universal Nonverbal Intelligence Test. Itasca, IL: Riverside.

Braden, J.P. (1999). Straight talk about assessment and diversity: What do we know. School Psychology Quarterly, 14, 343-351.

Brosman, F.L. (1983). Overrepresentation of low-socioeconomic minority students in special education programs in California. Learning Disability Quarterly, 6, 517-525.

Burns, R.B. (1994, April). Surveying the cognitive domain. Educational Researcher, 35-37.

Carroll, J.B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

Carroll, J.B. (1997). The three-stratum theory of cognitive abilities. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 122-130). New York: Guilford.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence. A critical experiment. Journal of Educational Psychology, 54, 1-22.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed) Hillsdale, NJ. Lawrence Earlbaum.

DeLeon, J. (1990). A model for an advocacy-oriented assessment process in the psychoeducational evaluation of culturally and linguistically different students. The Journal of Educational Issues of Language Minority Students, 7, 53-67.

Donovan, M.S. & Cross, C.T. (2002). Minority students in special and gifted education: Committee on minority representation in special education. Washington, DC: National Academy.

Elkind, D. (1975). Perceptual development in children. American Scientist, 63, 533-541.

Epstein, H.T. (1974a). Phrenoblysis: Special brain and mind growth periods: I. Human brain and skull development. Developmental Psychobiology, 7, 207-216.

Epstein, H.T. (1974b). Phrenoblysis: Special brain and mind growth periods: II. Human mental development. Developmental Psychobiology, 7, 217-224.

Eysenck, H.J. (1994). Personality and intelligence: Psychometric and experimental approaches. In R.J. Sternberg, P. Ruzgis, (Eds.), Personality and intelligence (pp. 3-31). New York, NY: Cambridge University Press.

Eysenck, H.J. (1998). A new look at intelligence. New Brunswick, NJ: Transaction Books.

Finlan, T.G. (1992). Do state methods of quantifying a severe discrepancy result in fewer students with learning disabilities? Learning Disability Quarterly, 15, 129-134.

Finlan, T.G. (1994). Learning disability: The imaginary disease. Westport, CT: Bergin & Garvey.

Flanagan, D.P., & Ortiz, S. (2001). Essentials of cross-battery assessment. New York: John Wiley & Sons.

Flynn, J.R. (1987). Massive gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101, 171-191.

Flynn, J.R. (1994). IQ gains over time. In R.J. Sternberg (Ed.), Encyclopedia of human intelligence (pp. 617-623). New York: Macmillan.

Flynn, J.R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), The rising curve: Long-term gains in IQ and related measures (pp. 25-66). Washington, DC: American Psychological Association.

Flynn, J.R. (1999). Searching for justice: The discovery of IQ gains over time. American Psychologist, 54, 5-20.

Frankenberger, W., & Fronzaglio, K. (1991). A review of states' criteria and procedures for identifying children with learning disabilities. Journal of Learning Disabilities, 23, 495-506.

Frisby, C.L. (1998). Culture and cultural differences. In J.H. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman, & J.R.Grenier (Eds.), Test interpretation and diversity: Achieving equity in assessment (pp. 51-73). Washington, DC: American Psychological Association.

Frisby, C.L. (1999). Culture and test session behavior: Part I. School Psychology Quarterly, 14, 263-280.

Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. New York: Basic Books.

Geisinger, K.F. (1998). Psychometric issues in test interpretation. In J.H. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman, & J.R.Grenier (Eds.), Test interpretation and diversity: Achieving equity in assessment (pp. 17-30). Washington, DC: American Psychological Association.

Glutting, J., & Oakland, T. (1993). Guide to the Assessment of Test Session Behaviors for the WISC-III and WIAT. San Antonio, TX: The Psychological Corporation.

Gould, S.J. (1981). The mismeasure of man. New York: Norton.

Gould, S.J. (1996). The mismeasure of man (Rev. ed.). New York: Norton.

Gustafsson, J.E. (2001). On the hierarchical structure of ability and personality. In J.M. Collis & S. Messick (Eds.), Intelligence and personality: Bridging the gap in theory and measurement (pp. 25-42). Mahwah, NJ: Erlbaum.

Gustafsson, J.E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28 (4), 407-434.

Herrnstein, R.J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.

Hilliard, A.G. (1992). The pitfalls and promises of special education practice. Exceptional Children, 59, 168-172.

Horn, J.L. (1991). Measurement of intellectual capabilities: A review of theory. In K.S. McGrew, J.K. Werder, & R.W. Woodcock, Woodcock-Johnson technical manual (pp. 197-232). Chicago: Riverside.

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology, 57, 253-270.

Horn, J.L., & Cattell, R.B. (1967). Age differences in fluid and crystallized intelligence. Acta Psychologica 26, 107-129.

Horn, J.L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 53-91). New York: Guilford.

Individuals With Disabilities Education Act. (1997). 1997 amendments [On-line]. Available: http://www.ed.gov/offices/osers/idea/the_law.html. - Mon Nov 27 12:01:44 EST 2000.

Ittenbach, R.F., Esters, I.G., & Wainer, H. (1997). The history of test development. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 17-31). New York: Guilford.

Jaynes, G.D. & Williams, R.M., Jr. (Eds.)(1989). A common destiny: Blacks and American society. Washington, DC: National Academy Press.

Jensen, A.R. (1974). Interaction of Level I and Level II abilities with race and socioeconomic status. Journal of Educational Psychology, 66, 99-111.

Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.

Jensen, A.R. (1998). The g factor : the science of mental ability. Westport, CT: Praeger.

Kamin, L. (1974). The science and politics of IQ. Hillsdale, NJ: Lawrence Erlbaum.

Kamphaus, R.W. (2001). Clinical assessment of child and adolescent intelligence. Needham Heights, MA: Allyn & Bacon.

Kamphaus, R.W., Petosky, M.D., Morgan, A.W. (1997). A history of intelligence test interpretation. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 32-47). New York: Guilford.

Kaufman, A.S., & Kaufman, N.L. (1983). Kaufman Assessment Battery for Children, Circle Pines, MN: American Guidance Service.

Keith, T.Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 373-402). New York: Guilford.

Keith, T.Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. School Psychology Quarterly, 14, 239-262.

Keith, T.Z., Kranzler, J. H., & Flanagan, D.P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock-Johnson Tests of Cognitive Ability-Third Edition (WJ-III). School Psychology Review, 30, 89-119.

Lambert, N.M. (1981). Psychological evidence in Larry P. v. Wilson Riles: An evaluation for the defense. American Psychologist, 36, 937-952.

Larry P. v. Riles, 343 F. Supp. 1306 (N.D. Cal. 1972, order granting preliminary injunction), aff'd 502 F. 2d 63 (9th Cir. 1974), 495 F. Supp. 926 (N.D. Cal. 1979, decision on merits), aff'd No. 80-427 (9th Cir. Jan. 23, 1984), No. C-71-2270 R.F.P. (Sept. 23, 1986, order modifying judgment).

Loehlin, J.C., Lindzey, G., & Spuhler, J.N. (1975). Race differences in intelligence. San Francisco: Freeman.

McGrew, K.S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 151-180). New York: Guilford.

McGrew, K.S., & Flanagan, D.P. (1998). The intelligence test desk reference: Gf-Gc Cross-battery assessment. Needham Heights, MA: Allyn & Bacon.

McGrew, K.S., & Woodcock, R.W. (2001). Technical Manual. Woodcock-Johnson III. Itasca, IL: Riverside Publishing.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Messick, S., & Anderson, S. (1970). Educational testing, individual development, and social responsiveness. Counseling Psychology, 2, 80-88.

Meyer, R.H. (1997). Value-added indicators of school performance: A primer. Economics of Education Review, 16, 283-301.

Morison, P., White, S.H., & Feuer, M.J. (Eds.). (1996). The use of IQ tests in special education decision making and planning. Washington, DC: National Academy Press.

Mosier, C.I. (1943). On the reliability of a weighted composite. Psychometrika, 8, 161-168.

Neisser, U. (Ed.), (1998). The rising curve: Long-term gains in IQ and related measures. Washington, DC: American Psychological Association.

Oakland, T. (Ed.) (1976). Non-biased assessment of minority group children: With bias toward none. Paper presented at a national planning conference on nondiscriminatory assessment for handicapped children. Lexington, KY.

Oakland, T., & Laosa, L.M. (1976). Professional, legislative, and judicial influences on psychoeducational assessment practices in schools. In T. Oakland (Ed.) (1976). Non-biased assessment of minority group children: With bias toward none. Paper presented at a national planning conference on nondiscriminatory assessment for handicapped children. Lexington, KY.

Ogbu, J.U. (1994). Culture and intelligence. In R. J. Sternberg (Ed.), Encyclopedia of human intelligence (Vol. 2, pp. 328-338). New York: Macmillan.

Onwuegbuzie, A.J., & Daley, C.E. (2001). Racial differences in IQ revisited: A synthesis of nearly a century of research. Journal of Black Psychology, 27, 209-220.

Opton, E. (1979). A psychologist takes a closer look at the recent landmark Larry P. Opinion. American Psychological Association Monitor, 10, (12), 1-4.

Ortiz, S.O. (2000). Best practices in nondiscriminatory assessment. Best Practices in School Psychology IV. Washington, DC: National Association of School Psychologists.

Parents in Action on Special Education v. Joseph P. Hannon, No. 74C 3586 (N.D. Ill.) (1980).

Plomin, R. (1988). The nature and nurture of cognitive abilities. In R.J. Sternberg (Ed.). Advances in the psychology of human intelligence. Vol. 4 (pp. 1-33). Hillsdale, NJ: Erlbaum.

Raven, J., Raven, J.C., & Court, J.H. (1993). Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 1). Oxford, England: Oxford Psychologists Press.

Reschly, D.J. (1981). Psychological testing in educational classification and placement. American Psychologists, 36, 1094-1102.

Reschly, D.J., & Ysseldyke, J.E. (1995). School psychology paradigm shift. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology (3rd ed.). Washington, DC: National Association of School Psychologists.

Reynolds, C.R., Lowe, P.A., & Saenz, A.L. (1999). The problem of bias in psychological assessment. In T. Gutkin & C. R. Reynolds (Eds.), The handbook of school psychology (3rd ed.). Washington, DC: National Association of School Psychologists.

Rushton, J.P. (1997). Race, intelligence, and the brain: The errors and omission of the "revised" edition of S.J. Gould's the mismeasure of man (1996). Personality and Individual Differences, 23, 169-180.

Salvia, J., & Ysseldyke, J. (1991). Assessment in special and remedial education (5th ed.), Boston: Houghton-Mifflin.

Sandoval, J.H. (1998). Critical thinking in test interpretation. In J.H. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman, & J.R.Grenier (Eds.), Test interpretation and diversity: Achieving equity in assessment (pp. 31-49). Washington, DC: American Psychological Association.

Sattler, J.M. (1988). Assessment of children (3rd ed.). San Diego: Author.

Sattler, J.M. (1992). Assessment of children's intelligence. In C.E. Walker, & M.C., Roberts (Eds.), Handbook of clinical child psychology (2nd ed., pp. 85-100). New York, NY: John Wiley & Sons.

Sattler, J.M. (2001). Assessment of children: Cognitive applications (4th ed.). San Diego: Author.

Scarr, S. (1978). From evolution to Larry P., or what shall we do about IQ tests? Intelligence, 2, 325-342.

Scheuneman & Oakland, T. (1998). High-stakes testing in education. In J. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman, & J.R.Grenier (Eds.), Test interpretation and diversity: Achieving equity in assessment (pp. 77-103). Washington, DC: American Psychological Association.

Spearman, C.E. (1923). The nature of intelligence and the principles of cognition. London: Macmillan.

Spearman, C.E. (1927). The abilities of man. New York: Macmillan.

Sternberg, R.J. (1994). A triarchic model for teaching and assessing students in general psychology. General Psychologist, 30 (2), 42-48.

Styles, I. (1999). The study of intelligence--The interplay between theory and measurement. In M. Anderson, (Ed.), The development of intelligence. Studies in developmental psychology (pp 19-42). Hove, England: Psychology Press/Taylor & Francis.

Taylor, O.L. (1989). Clinical practice as a social occasion. In L. Cole & V. Deal (Eds.), Communication Disorders in Multicultural populations (pp. 18-27). Rockville, MD: American Speech-Language Hearing Association.

Thurstone, L.L. (1938). Primary mental abilities. Psychometric Monographs (1).

Thurstone, L.L., & Thurstone, T.G. (1941). Factorial studies of intelligence: Psychometric Monographs, No. 2.

Twenty-second Annual Report to Congress on the Implementation of the Individuals With Disabilities Education Act. (2000). [On-line]. Available: http://www.ed.gov/offices/OSERS/OSEP/Products/OSEP2000AnlRpt/PDF/Chapter-2.pdf - 415143 bytes - Mon Nov 27 12:01:44 EST 2000.

U.S. Bureau of the Census. (2000). Racial population estimates, (January, 2001). Washington, DC. Government Printing Office.

U.S. Department of Education, Office for Civil Rights. (2000). The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers.

U.S. Department of Education, Office for Civil Rights. (1997). Fall 1994 elementary and secondary school civil rights compliance report. Washington, DC: Author.

Urbach, P. (1974). Progress and degeneration in the "IQ debate." British Journal of the Philosophy of Science, 25, 99-135, 235-259.

Valencia, R.R., & Suzuki, L.A. (2001). Intelligence testing and minority students: Foundations, performance factors, and assessment issues. Thousand Oaks, CA: Sage.

Wesman, A.G. (1968). Intelligent testing. American Psychologist, 23, 267-274.

Wesson, K. A. (2000). The Volvo effect – Questioning standardized tests. Education Week, 20, 34-36.

Westby, C. (1990). There's no such thing as culture-free testing. Texas Journal of Audiology and Speech Pathology, Spring/Summer, 4-5.

Ysseldyke, J.E., Algozzine, B., & McGue, M. (1995). Differentiating low-achieving students: Thoughts on setting the record straight. Learning Disabilities Research & Practice, 10, 140-144.

BIOGRAPHICAL SKETCH

Oliver W. Edwards completed his undergraduate studies in psychology at Florida International University in 1986. He completed two graduate degrees in school psychology at the University of Florida in 1989. After graduating from the University of Florida, he practiced as a school psychologist with the School Board of Broward County, Florida. As a staff psychologist, his role included instruction, assessment, consultation, intervention development/implementation, and counseling students and families about every issue that could impact the students' school functioning. He later became an administrator with the district, supervising roughly 65 school psychologists and school social workers in their work with 65 schools and some 75,000 students. As an administrator, he worked with superintendents, principals, parents, and teachers regarding student services issues. Although he has published in a refereed educational law journal on special education law topics, his current research interests focus on theories of intelligence and the sociology of education. He has published several papers in peer-reviewed journals and was also invited to write a book chapter about the latter topic. Currently, he is researching issues involving utilizing family and social support networks to aid students' academic and emotional functioning. He also has a strong interest in high-stakes testing and intends to conduct research in this area.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas D. Oakland, Chair
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Nancy Waldron
Associate Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

M. David Miller
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

W. Max Parker
Professor of Counselor Education

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 2003

Dean, College of Education

Dean, Graduate School